UNIVERSITY OF CALIFORNIA
RIVERSIDE

Functional Characterization of the Active Muta1 Transposable Element and Pack-Mutator-Like Elements in the Mosquito Aedes aegypti

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Plant Biology

by

Kun Liu

March 2017

Dissertation Committee:
Dr. Susan Wessler, Chairperson
Dr. Peter Atkinson
Dr. Jason Stajich

The Dissertation of Kun Liu is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Susan Wessler, from whom I learnt so much. She is a brilliant scientist and a gracious mentor. Without her constant guidance, support and encouragement, this work would not have been possible. I am also deeply grateful to Dr. Peter Atkinson and Dr. Jason Stajich, who provided insightful suggestions with great patience. My gratitude also goes to the members of my Qualifying Exam Committee, Dr. Xuemei Chen, Dr. Linda Walling, and Dr. Zhenbiao Yang for their guidance and critical advice.

I wish to thank all the members of the Wessler lab for their help and support and for all the great time we spent together that will be missed. My special thanks go to Dr. James Burnett for teaching and training me, for giving me directions and inspiration at the beginning of my PhD study. I would also like to thank Lu Lu, Jinfeng Chen, Jinhua Shi and Brad Cavinder for sharing their knowledge and expertise, for helping me with difficulties in my research. I would like to thank Alex Cortez, Venkateswari Jaganatha Chetty, Tingting Ma, Sofia Rob, Matt Collin and Courun Williams for their friendship, for the great discussions and for their constant help.

I have also got a lot of help from people outside the Wessler lab. I thank Dr. Rob Hice and Presha Shah for the fruitful collaboration in the *Muta1* transposition assays. I wish to thank the entire Department of Botany and Plant Sciences and the Institute

for Integrative Genome Biology (IIGB) at UC Riverside for providing a most pleasant and inspiring academic environment.

Finally, my deepest gratitude goes to my parents and my wife for being there, for caring, for their patience and love that supported me throughout these years.

ABSTRACT OF THE DISSERTATION

Functional Characterization of the Active Muta1 Transposable Element and Pack-Mutator-Like Elements in the Mosquito Aedes aegypti

by

Kun Liu

Doctor of Philosophy, Graduate Program in Plant Biology
University of California, Riverside, March 2017
Dr. Susan Wessler, Chairperson

Mutator–like transposable elements (MULEs) are widespread across fungi, plants and animals. Most of the research of MULEs has focused on plant where they are discovered and have significant impact on genome structure. Despite being widespread, only a few active MULEs have been identified, meanwhile, the transposition mechanism of the MULEs is previously unknown. Pack-MULEs are able to capture and amplify genes or gene fragments on a large scale, and a subset of plant Pack-MULEs have been shown to be likely playing functional roles in regulating gene expression and providing novel coding capacities. However, Pack-MULEs have only been reported in plant species.

We report that *Muta1* identified from the mosquito *Aedes aegypti* is capable of excision and reinsertion in a yeast transposition assay, element reinsertion generated either 8 bp or 9 bp target site duplications (TSDs) with no apparent sequence preference. Mutagenesis analysis revealed the importance of several

conserved amino acids, including the DDE triad in transposase function; donor site TSD also impacts the transposition of *Muta1*. Via *in vitro* assays, we have dissected the process of DNA breakage and end joining of *Muta1*. The transposition reactions involve double strand break with hairpin formation on the flanking DNA, and 3' OH joining to the target DNA, which is also observed in the *hAT* transposons and the V(D)J recombination system. Mutagenesis analysis revealed the involvement of the DDE triad and a conserved W residue in transposition reactions. The terminal motif and subterminal repeats also impact *Muta1* transposition. We also report the identification 1,378 Pack-MULEs from the *A. aegypti,* genome. Pack-MULEs preferentially acquire genic fragments; they contain fragments from 423 parental genes. Of the *A. aegypti* Pack-MULES, 9.7% are expressed, 11.8% directly generate sRNAs, and no evidence of the translation of Pack-MULE sequences could be detected. Comparison of the sequences acquired by the host genes indicates that fragments of genomic DNA have been captured, amplified and rearranged on a large scale and for a long period of time in the *A. aegypti* genome. Gene acquisition activity of Pack-MULEs might provide an important mechanism for the evolution of genes in *A. aegypti*.

Table of Contents

List of Figures

List of Tables

Chapter 1 Functional characterization of the active *Mutator*-like transposable element, *Muta1* from the mosquito *Aedes aegypti*

Abstract

Mutator-like transposable elements (MULEs) are widespread with members in fungi, plants, and animals. Most of the research on the MULE superfamily has focused on plant MULEs where they were discovered and where some are extremely active and have significant impact on genome structure. The maize *MuDR* element has been widely used as a tool for both forward and reverse genetic studies because of its high transposition rate and preference for targeting genic regions. However, despite being widespread, only a few active MULEs have been identified, and only one, the rice *Os3378*, has demonstrated activity in a non-host organism.

Here we report the identification of potentially active MULEs in the mosquito *Aedes aegypti.* We demonstrate that one of these, *Muta1*, is capable of excision and reinsertion in a yeast transposition assay. Element reinsertion generated either 8 bp or 9 bp target site duplications (TSDs) with no apparent sequence preference. Mutagenesis analysis of donor site TSDs in the yeast assay indicates that their presence is important for precise excision and enhanced transposition. Site directed mutagenesis of the putative DDE catalytic motif and other conserved residues in the transposase protein abolished transposition activity.

Introduction

Transposable elements (TEs) are mobile fragments of DNA that can move from one locus to another in the host genome, often replicating in the process. TEs usually make up the largest fraction of eukaryotic genomes; accounting for almost half of the human genome, and more than 70% of the genomes of some grass species (Finnegan, 1989). Based on their transposition intermediate, eukaryotic TEs are divided into two classes. Class 1 elements utilize an RNA intermediate in the transposition reaction while the intermediate for most class 2 elements is the element itself that is mobilized by a 'cut and paste' mechanism (Wicker et al., 2007). A TE family is composed of one or more transposase-encoding autonomous elements and up to several thousand nonautonomous elements that do not encode functional transposase. Family members share the same terminal inverted repeats (TIRs) and target site duplications (TSDs) of the same length allowing them to move in *trans* by utilizing the transposase encoded by the autonomous family member (Feschotte et al., 2007).

Prior studies have classified class 2 TEs into upwards of 19 superfamilies on the basis of the relatedness of the element-encoded transposase (Bao et al., 2007). The original *Mutator* element, now called *MuDR*, was first isolated from a maize strain as the agent responsible for its high forward mutation rate (Robertson 1978; Lisch, 2013). Subsequently, members of this superfamily, called collectively *Mutator*-like transposable elements (MULEs), were found in other plants, and in fungi,

protozoans, and in a variety of animals, (from insects to fish to metazoans, Neuveglise et al., 2005; Pritham et at., 2005; Lopes et al., 2009). Typical features of MULEs include long terminal inverted repeats (TIRs) (>100 bp) and an 8-10 bp TSD (Lisch 2002). Nonautonomous family members often contain a variety of sequences between the TIRs, including fragments from host genes; such elements are called Pack-MULEs (Jiang et al., 2004). To date, only a few active MULEs have been identified, including *MuDR*, *Hop1*, Jittery, *TED*, and *Os3378* (Lisch 2013; Chalvet et al., 2004; Xu et al., 2004; Li et al., 2013; Zhao et al., 2015). Importantly, only the rice *Os3378* element has been shown to transpose in a heterologous host (Zhao et al., 2015).

Most MULE transposases contain a N-terminal zinc finger DNA binding motif and a conserved C-terminal DDE domain, which has been shown to be the catalytic core for transposition reactions in other superfamilies but not to date for MULEs (Babu et al., 2006; Nesmelova et al., 2010). Phylogenetic studies indicate that the DDE domain of MULE transposases is closely related to the IS*256* family, which is present in diverse prokaryotes (Eisen et al., 1994). Each residue of the predicted DDE motif of IS*256* has been shown to be necessary for transposition (Loessner et al., 2002). Most MULE transposases also harbor a CH motif and a W residue between the second D and E of the DDE domain, which are also conserved in the *hAT* superfamily. Mutagenesis analyses in *hAT* elements demonstrated the importance of these residues for transposition (Yuan et al., 2011; Hickman et al., 2014). Whether

these residues are also important for the MULE transposase has not as yet been determined.

TE-based genetic tools have facilitated our deep understanding of the biology of both plants and animals. One of the most important species lacking such tools is the mosquito, *A. aegypti,* which can spread dengue fever, yellow fever, chikungunya, zika, and many other diseases that are responsible for over one million deaths annually (Womack M 1993; Marchette et al., 1969). Four exogenous TEs that transpose at high frequencies in *Drosophila melanogaster* (*Hermes, Mos1, Minos*, *piggyBac)* were found to move rarely or not at all in the germ line of *A. aegypti* as very few germinal mutations were detected (O'Brochta et al., 2003; Fraser et al., 2012). To explain this result, it was hypothesized that mosquitoes have a strong genome defense system that could effectively recognize and silence foreign TEs (Scali et al., 2007). Therefore endogenous active TEs could be effective mutagens from generation to generation as they might be able to evade genome surveillance. Nearly 20% of the 1.4 Gb genome of *A. aegypti*, is derived from class 2 TEs representing 17 superfamilies, which enabled the search for potentially active TEs in the genome (Nene et al., 2007). In this study we show that *Muta1,* identified by computer-assisted analysis of the *A. aegypti* genome, is an active TE capable of both insertion and excision in a yeast transposition assay. Site-directed mutagenesis analysis revealed that transposition activity in yeast was influenced by disruption of several conserved residues and by the presence of TSDs at the donor site. With

characteristics such as high transposition activity, precise excision, and no target sequence preference, *Muta1* could be crafted into an effective tool for forward mutation analysis in mosquitoes.

Results

Characterization of the MULE superfamily in *Ae.*

Over 1000 MULE*s*, including 3 full-length elements, were previously reported in the *A. aegypti* genome (Marquez et al., 2010). To identify potentially active MULEs, sequences of conserved DDE domains were used as queries in TBlastN searches against the *A. aegypti* genome through the TARGeT program, which is designed for TE discovery (Yuan et al., 2011; Han et al., 2009). After removal of duplicate hits, 141 sequences with significant similarity (e- value $<10^{-15}$) to the MULE DDE domain were identified. Full-length elements were defined by comparison of sequences in the same branch of the phylogenetic tree where adjacent sequences share high similarity that extends beyond the DDE domain. Sequences near the boundary of similarity were examined manually for the long TIRs and 8-10 bp TSDs that are features of MULEs.

Using this methodology, 31 full-length elements sorted into 14 families, each with > 95% sequence similarity (*Muta1-14,* Figure 1-1). All 14 families have 8-9 bp TSD, 12 of 14 have TIRs >100 bp and two have TIRs <60 bp. Ten of the 12 long TIR families contain short subterminal tandem repeats (9-15 bp). All families except *Muta6* also contain derivative nonautonomous elements, which share high sequence similarity in their TIR and subterminal regions, but carry heterogeneous internal sequences including fragments of host genes.

The phylogenetic tree also reveals the evolutionary relationships of the 14 families. Among the 5 major groups resolved on the tree, group B includes half of the 14 families, while group D and E contains one family each (Figure 1-1). Furthermore, long branches indicate extensive sequence differences between families. For example, although *Muta5* and *Muta8* belong to group A and locate to adjacent clades, no significant nucleotide similarity was detected between these elements. In another example, nucleotide similarity between *Muta3* and *Muta11* is restricted to the DDE region (71% identity, ~390 bp) and the TIRs (76% identity, ~ 110 bp).

After removal of other TE insertions and manual correction of frameshifts caused by small insertions and deletions, the 31 full-length elements were predicted to encode transposase proteins ranging in size from 416 to 554 residues. Comparison of conceptually translated transposases identified conserved regions, other than the DDE domain, that could have functional significance. The transposases of 13 families were predicted to harbor a FLYWCH type zinc finger DNA binding domain in the N-terminus, while *Muta6* was predicted to harbor a SWIM type zinc finger DNA binding domain in the C-terminus. Each family contains at least one putative full-length copy in which the coding region is not interrupted by frameshifts or stop codons. Of particular interest to this study, 3 families (*Muta1, Muta3* and *Muta5*) include at least 2 members that are identical or nearly identical (>99% identity).

These features indicate recent and possible ongoing activity of multiple MULE families in *A. aegypti*.

Identification of the active *Muta1* family

Of the 14 families, *Muta1* appeared to be the best candidate for an active element. The family contains 7 identical copies and an 8th copy with only two point mutations in predicted noncoding DNA. *Muta1* is 3198 bp, flanked by TSDs of 8 bp or 9 bp, TIRs of 145 bp comprised of a 10 bp imperfect palindromic terminal motif with the 5th and 6th nucleotide unpaired, and 9 copies of a 12 bp subterminal tandem repeat separated by 3-4 bp spacers (Figure 1-2, open and solid arrows, respectively). Because of the complexity of the TIR structure, *Muta1* can be classified as a type2 Foldback TE (Bingham et al., 1989). The *Muta1* transposase is predicted to be 504-amino acid transposase and encoded by two exons.

Over 300 putative nonautonomous elements derived from *Muta1* were detected in the *A. aegypti* genome (e- value <10^{-10}), with 171 flanked by perfect TSDs of 8 bp or 9 bp (Figure 1-3A). Most of the derivative elements share the TIR sequence with *Muta1,* but the TIRs are often truncated, with variable copies of the subterminal repeats (Figure 1-4). There are multiple copies of some derivative elements; for example, there are 4 copies of *Muta1NA1* (>98% identical). Most derivative elements contain variable internal sequences and share sequence similarity with only the TIR of *Muta1*. For about 20% of the derivative elements these variable

internal regions can be aligned with sequences from host genes, much like previously described Pack-MULEs (Jiang et al., 2004). For example, the 1623 bp *Muta1NA3* contains a 276 bp fragment from a serine/threonine-protein kinase gene (97.5% identical, e-value< 1e-51) (Figure 1-4). Among the 171 insertion sites, 110 (64.3%) are located in gene bodies or within 5kb upstream or downstream of genes. In a control dataset of 171 randomly selected genomic sites (see methods) only 49 (28.7%) were located in the same regions (Figure 1-3B), suggesting that *Muta1* may have an insertion preference for genic regions as was previously reported for plant MULEs (Lisch 2003).

Muta1 can transpose in yeast

A yeast transposition assay was employed to determine whether *Muta1* transposase is able to catalyze the movement of natural and/or artificial nonautonomous elements. In prior studies members of five superfamilies (MULE, *Tc1/mariner, hAT, PIF/Harbinger* and *piggyBac*) were shown to transpose in yeast (Zhao et al., 2015; Yang et al., 2004; Weil et al., 2000; Hancock et al., 2010; Mitra et al., 2008). Our yeast assay consisted of two plasmid constructs: an expression vector containing the *Muta1* transposase coding sequence downstream of the galactose inducible *GAL1* promoter, and a reporter vector containing a nonautonomous element inserted in, and blocking expression of, the *ADE2* gene (Figure 1-5A). We first tested the artificial *Muta1AR* element, which contains 250 bp from each end of *Muta1* (Figure 3). Transposase mediated TE excision restored *ADE2* expression and resulted in

reversion of adenine auxotrophy to prototrophy (Figure 1-5B). Excision events were validated by PCR amplification of the *ADE2* empty site from revertants (Figure 1-6A).

In subsequent experiments, the reporter plasmid was modified by inserting several natural nonautonomous elements (*Muta1NA1-5*, Figure 1-4) into the *ADE2* coding region. Despite differences in TIR length and the length and sequence of internal regions, *Muta1* transposase was able to mobilize all of these elements, but none with a frequency as high as the artificial element *Muta1AR*.

Integration events catalyzed by *Muta1* transposase were assayed by first constructing the nonautonomous *Muta1HIS* element containing a yeast selectable marker *HIS3* gene flanked by 250 bp of the *Muta1* termini (Figure 1-4). Integration of *Muta1HIS* into yeast chromosomes was assayed in plasmid-free cells following selection with 5-FOA, (see Methods, Figure 1-5D). Comparison of the frequency of *ADE2* revertants retaining the *HIS3* marker and are 5-FOA resistant (2.71×10^{-6} His^+ $5\text{-}FOA^R$ cells/total cells) to the frequency of total *ADE2* revertants (2.44×10^{-5} Ade^+ cells/total cells) indicated that about 11% of the *Muta1HIS* elements excised from the donor plasmid had reintegrated in yeast chromosomes (Table 1-2). In another assay, *ADE2* colonies isolated directly from the *Muta1HIS* excision assay were tested to determine if they were also His^+ and $5\text{-}FOA^R$. Of 300 revertant colonies, 41 (14%)

could proliferate on selective plates, in agreement with the results of the first approach (Table 1-1).

To determine the precise insertion sites, polymorphic bands on transposon display gels were recovered (Biedler et al., 2003), sequenced and mapped to yeast chromosomes (Figure 1-6B). Sixty of 62 sites had significant matches with yeast genomic sequence, while two matched plasmid sequences. Fourteen of the 60 genomic sites were in gene bodies while 27 were within 1kb of genes. Thus 41/60 insertions (68%) were in gene-rich regions (Lynch et al., 2008). Amplification and sequencing of each site revealed the presence of TSDs with 8 bp for 21 events and 9 bp for 39 events. Consensus sequences generated from the yeast insertion sites and the 171 sites for *Muta1* derivative elements in the *A. aegypti* genome indicated little or no sequence preference for insertion (Figure 1-7).

The *Muta5* element does not transpose in yeast

The successful transposition of *Muta1* in yeast prompted us to perform a similar analysis of the *Muta5* family which contains 3 identical copies, a 4th with >99% sequence identity, and 3 copies with large deletions or insertions (Figure 1-1). *Muta5* is 3496 bp, flanked by TIRs of 151 bp and TSDs of 8 bp or 9 bp and is predicted to contain 3 exons that encode a 554-amino acid transposase. The TIR of *Muta5* contains 9 copies of a 15 bp subterminal tandem repeat and an 8bp terminal motif (Figure 1-8A). There are over 200 *Muta5* derivative elements (e- value <10^{-15})

in the *A. aegypti* genome, with 93 flanked by 8 bp or 9 bp TSD. Taken together the features of the *Muta5* family strongly suggested that it was an active element. However, in the yeast transposition assay the *Muta5* transposase was unable to catalyze transposition of *Muta5AR* (Figure 1-8A), which contains 250 bp from the ends of *Muta5* (Figure 1-8B).

TSD at donor site affects *Muta1* transposition in yeast

Successful transposition of *Muta1* in yeast facilitated the analysis of the importance of its features by quantifying the impact of mutations on transposition quality and frequency. With regard to the role of the TSDs, although the nonautonomous constructs used in the assays described thus far lacked flanking TSDs, their reinsertion still generated TSDs of 8 bp or 9 bp. To examine the impact of TSD length or sequence on excision frequency of *Muta1AR*, three versions of 8 bp (TTCAATAG, CGATTCAA and GGTAACTC) or 9 bp (ATTCAATAG, TCGATTCAA and CGGTAACTC) TSDs were tested. Addition of 8 bp or 9 bp TSDs at the donor site increased *Muta1AR* excision frequency by ~7 fold and ~3 fold, respectively when compared to the controls lacking TSDs (Table 1-2). Similarly, introduction of TSDs flanking *Muta1HIS* increased reintegration by about 40% and 90% for 8 bp TSDs or 9 bp TSDs respectively (Table 1-2). For both excision and integration, TSD sequence had little impact.

We next addressed the question of whether the presence or absence of TSDs at the donor site impacts the quality of excision by analyzing so-called transposon footprints. Specifically, class 2 TEs often leaves a footprint upon excision consisting of a few nucleotides or small rearrangements at the site of excision site (Sutton et al., 1983). Formation of footprints involves DNA repair of sequences flanking the excised element. To assess the impact of TSDs on the repair of excision sites, the donor element construct was modified so that all excision footprints (not only those that maintain the reading frame) could be analyzed. First, nonautonomous elements were inserted in the 5' UTR of *ADE2* (Figure 1-5C). Second, because insertion of *Muta1AR* in the 5'UTR resulted in leaky *ADE2* expression, we substituted the longer *Muta1NA1* (Figure 1-4), which, in the absence of transposase, blocked *ADE2* expression. When either 8 bp or 9 bp TSDs were added to *Muta1NA1*, about 90% of revertants were precise (Table 1-2), meaning that the element was removed as well as a single copy of the TSD. The quality of excision appeared to be independent of TSD sequence (Figure 1-9B-G). In contrast, absence of donor site TSDs reduced perfect excision to only 10% of the *ADE2* revertant colonies sequenced (Table 1-2). Most excision sites reflected loss of a few nucleotides from either side of the flanking DNA. Occasionally, part of the TIR (up to 13 bp) was left after excision and repair (Figure 1-9A).

Mutagenesis of *Muta1* transposase

The catalytic domains of all characterized transposases of class 2 TEs contain a DDE/D amino acid triad (Yuan et al., 2011) and mutagenesis studies confirmed its functional significance in the *piggyBac, Mariner* and *hAT* superfamilies (Mitra et al., 2008; Brillet et al., 2007; Zhou et al., 2004). Alignment with the transposase of other active MULE*s* showed that the DDE triad in *Muta1* corresponds to D214, D283 and E419 (Figure 1-10A). To determine if these conserved sites play key roles in transposition, site-directed mutagenesis was performed. Transposition activity was completely abolished when D214, D283 or E419 was mutated to alanine (Figure 1-10B). In contrast, mutation of nonconserved sites, including E129, E188, E239, W313 and D473 to alanine had little impact. Although E373 is not a conserved site, mutation to alanine also completely abolished transposition activity.

The functional significance of two additional highly conserved residues in *Muta1*, H307 and W401, were tested. In a prior study these residues were shown to be essential for transposition of the *hAT* superfamily member *Hermes* (Zhou et al., 2004; Hickman et al., 2014). Mutation of the corresponding *Hermes* H and W residues (H268 and W319) to alanine completely abolished transposition activity. Similarly for the *Muta1* transposase, mutation of H307 or W401 to alanine abolished activity (Figure 1-10B). Analysis of *Hermes* transposase also showed that the W319 residue was likely necessary for the correct positioning of flanking DNA during the excision reaction, and that other aromatic residues can partially substitute for this

function (Hickman et al., 2014). When W401 of *Muta1* was mutated to phenylalanine, transposition activity was reduced by 79% and the frequency of precise excision dropped from 90% to approximately 48% (8 bp donor site TSD, 14/29 events) and 50% (9 bp donor site TSD, 15/30 events) (Figure 1-9H&I), suggesting that the *Muta1* W401F mutation also led to inaccurate positioning. These results confirmed the importance of the putative DDE motif, the conserved H307, W401 and identified the nonconserved E373 as a potentially important residue for transposase function.

Discussion

The MULE superfamily is widespread in eukaryotic genomes and is closely related to prokaryotic IS*256* elements. However, although it is also found in the genomes of many insects, no active elements have been reported in insect species. In this study, we performed a thorough search for potentially active MULEs in the *A. aegypti* genome and demonstrate that *Muta1* encodes a transposase that catalyzes the excision and reinsertion of nonautonomous derivative elements in yeast. With the availability of this heterologous transposition assay, the function of the conserved MULE DDE domain and the role of TSD in transposition were tested.

The DDE/D domain is proposed to be the catalytic core involved in transposition of class 2 TEs and has been identified in all superfamilies (Yuan et al., 2011). Prior to this study, the functional significance of this domain had been experimentally validated only for members of the *piggyBac*, *Tc1/Mariner*, and *hAT* superfamilies (Mitra et al., 2008; Brillet et al., 2007; Zhou et al., 2004). Results of this study provide the first experimental evidence for the importance of the DDE motif in the transposition reaction of *Muta1*, a member of the MULE superfamily. Specifically, transposition was completely abolished when any of the three residues were mutated to alanine (Figure 1-10B).

In addition to the DDE triad, other residues critical for transposition were identified. W401 of *Muta1* is a conserved residue that is also found in the *hAT* superfamily

(Yuan et al., 2011; Zhou et al., 2004). Crystallographic analysis and *in vitro* biochemical assays showed that the corresponding W318 residue in *Hermes* functions in the positioning of flanking DNA, which ensures that the double strand break occurs at the correct position when an element excises from flanking DNA (Hickman et al., 2014). Other aromatic residues partially substitute for its function, however the mutant transposase generated additional species of intermediates in double strand break repair than the wild-type transposase, suggesting that inaccuracy in the position of the cleavage site may be the cause by these mutations (Hickman et al., 2014). For *Muta1,* the W401A mutation completely abolished transposition activity while the W401F mutation resulted in a 79% reduction of transposition frequency (Figure 1-10B) and caused inaccurate excision as the frequency of precise excision dropped from 90% to ~50% (Figure 1-9 H&I). Taken together these data suggest that this conserved tryptophan residue is likely playing a similar role in *hAT* and MULE transposases, which is to correctly position flanking DNA for the excision reaction. In addition to the W residue, a CxxH motif is also shared between MULE and *hAT* elements, the *Hermes* H268 was found to be located close to the DDE active center and involved in the interaction with TIRs (Hickman et al., 2014). Mutation of the corresponding H307 to alanine resulted in a 99.5% reduction of *Muta1* transposition activity (Figure 1-10B), suggesting the importance of this residue for *Muta1* transposase function. Taken together, our study provides experimental evidence to support the close evolutional relationship reported previously between the MULE and *hAT* superfamilies (Yuan et al., 2011).

Prior to this study, the only MULE shown to transpose in a heterologous host was the rice *Os3378* element (Zhao et al., 2015). Because both *Muta1* and *Os3378* have demonstrated activity in yeast assays that employed very similar experimental design and nonautonomous elements of similar size, comparison of assay results may be informative (Table 1-3). Excision frequencies of the 500 bp *Muta1AR*, as high as 6940 events per 10^7 cells (Figure 1-2), is ~320 fold higher than *Os3378NA469* (469bp). For both elements, excision frequencies are increased by the presence of donor site TSD with *Muta1AR* enhanced by ~7 fold (8 bp) and 3 fold (9 bp), and *Os3378NA469* enhanced by ~17 fold with TSDs of 9 bp (Table 1-3). About 80% of *Os3378* integration sites in the yeast genome were located in gene bodies or within 1 kb of flanking regions of genes while *Muta1* had a slightly lower ratio of 68%. In summary *Muta1* shows very similar transposition behavior as *Os3378,* and the higher activity of *Muta1* makes it a better tool for the future study of MULE transposition, for example, the biochemical process of excision and integration, and how Pack-MULEs capture host gene fragments (Jiang et al., 2004).

The presence of donor site TSDs impacts the quality of *Muta1*-mediated excision events as various footprints were generated without donor TSD (Figure 1-9A). The predominance of small deletions (1-4 bp) suggests that the *Muta1* transposase cuts outside the TIR. In contrast, with the presence of either 8 bp or 9 bp TSD at the donor site, most excision events were precise and the actual TSD sequence did not seem to matter (Figure 1-9 B-G). Similar behavior was also observed for IS*256*, the

prokaryotic TE family related to MULEs, and for the one other MULE tested, Os3378. Reduction of TSD from 8bp to 6bp eliminated precise excision of IS*256* and reduced the *Os3378* precise excision frequency from 97.44% to 82.05% (Zhao et al., 2015; Loessner et al., 2002). For IS*256* it was hypothesized that precise excision is achieved through a transposase-independent replication slippage mechanism that requires a short stretch of homologous DNA with a minimum length of 8 bp (Hennig et al., 2008). In our assay, the absence of donor TSDs resulted in a 90% reduction in precise excision (Figure S1-9 A-G), which suggests a role for TSDs in promoting precise excision.

Thirty-one full-length MULEs that group into 14 families were identified in the *A. aegypti* genome. Several families have identical or nearly identical full-length copies including *Muta1* (7 identical), *Muta5* (3 identical), and *Muta3* (2 with only 2 noncoding SNPs). Although the existence of identical genomic copies is a feature of active TEs, *Muta5* was unable to catalyze the movement of nonautonomous derivative elements in yeast (Figure 1-8). One explanation for our success with *Muta1* but failure with *Muta5* is that the latter has 3 predicted exons while the former has 2. More predicted exons would increase the chances of incorrectly assembling the actual/functional *Muta5* transposase.

Accumulation of seven identical copies of *Muta1* in *A. aegypti* suggests that this element may still be active or that it has some success evading the genome surveillance system shown previously to effectively silence exogenous TEs (O'Brochta et al., 2003; Scali et al., 2007). In this regard, it may be possible to engineer *Muta1* to make it an effective endogenous mutagen. Like the *MuDR* system in maize, where the genome has numerous copies of nonautonomous *Mu* elements, there are over 300 *Muta1* derivative elements in the *A. aegypti* genome (Figure 1-3A). If *Muta1* was able to mobilize even a subset of these elements as it does in yeast, it could be an effective tool for high frequency insertional mutagenesis, especially when coupled with its preference for genic insertions and a lack of target sequence preference (Figure 1-3B & 1-7).

Conclusions

This is the first report of the transposition of a non-plant MULE, *Muta1,* in a heterologous system and provides the first experimental evidence for the functional significance of the DDE domain in the transposition reaction in the MULE superfamily. High frequency transposition in a yeast assay facilitated the determination of *Muta1* transposition features including precise excision, genic targeting with no sequence preference and the impact of TIR and TSD for insertion and excision. Taken together, *Muta1* may be a valuable tool for forward genetics in mosquitoes.

Material and Methods

Identification of MULEs in *A. aegypti*

The conserved MULE DDE domain from all eukaryotes (Yuan et al., 2011) was used as query to search the *A. aegypti* genome (AaegL3 build, https://www.vectorbase.org/organisms/aedes-aegypti/liverpool/aaegl3) by TBLASTN, as implemented in the TARGeT pipeline (Han et al., 2009) with an E-value cutoff of 0.001. Flanking DNA sequences with 10 kb upstream and downstream of the matched region were retrieved. The ends of a putative element were determined by aligning two closely related elements with their 20 kb flanking sequences, TIR boundaries and TSDs were manually identified. Coding capacity of each element was predicted by the GENSCAN program (http://genes.mit.edu/GENSCANinfo.html).

To identify *Muta1* derivative nonautonomous elements, 50 bp from each end of *Muta1* was used in a BLASTN search with TARGeT (Han et al., 2009) using default parameters. One hundred bp of flanking DNA sequences were retrieved for manual verification of the TIR and TSD of each derivative element. Fifty bp flanking each element were used for BLASTN searches against the *A. aegypti* genome (AaegL3 build) to determine the genomic location and compared to the genome annotation (release AaegL3.3, https://www.vectorbase.org/organisms/aedes-aegypti/liverpool/aaegl3) to determine the adjacent genes. For the control data set, 171 genome coordinates across the 4,757 scaffolds were randomly generated, and

compared to the genome annotation (release AaegL3.3) to determine the surrounding sequences and genes. The random insertion sites generation used 1,000 replicates to estimate the expected number of insertions (and standard deviations) in each category.

Yeast construct construction

Genomic DNA of individual *A. aegypti* mosquito (Liverpool strain, obtained from Dr. Atkinson, UC Riverside) was extracted using the DNeasy Blood & Tissue Kit (Qiagen). The two exons of *Muta1* predicted by GENSCAN program were cloned from genomic DNA and fused through overlap PCR. The complete transposase coding sequence was then cloned into the Gateway cloning vector pENTR and transferred to destination vector pGAL415-ccdb (Alberti et al., 2007) with LR Clonase (Invitrogen) to generate the pGAL415-ccdbMuta1 plasmid (Figure 1-5A).

Two hundred fifty bp from each end of *Muta1* were fused by overlap PCR to generate *Muta1AR.* The *HIS3* fragment containing the yeast *HIS3* coding sequence, *HIS3* 5' and 3' UTR, and *HIS3* promoter was cloned from vector pGAL415-ccdb (Alberti et al., 2007). The *HIS3* fragment was then fused with 250 bp from each end of *Muta1* through overlap PCR to generate *Muta1HIS. Muta1NA1-5* elements were cloned directly from genomic DNA. All nonautonomous elements were inserted in the *Hpa*I site of *ADE2* for the exon excision assay or the *Xho*I site for 5' UTR excision assay through homologous recombination in yeast as previously described

(Hancock et al., 2010). Donor site TSDs were introduced by adding corresponding TSD sequences in primers.

For *Muta5* assay, plasmid pGAL415-ccdbMuta5 was constructed in the same way as pGAL415-ccdbMuta1, and *Muta5NA* was constructed by overlap PCR.

Transposition assay

The yeast transposition assay using *Saccharomyces cerevisiae* strain DG2523 and the pWL89a vector was described previously (Yang et al., 2006; Weil et al., 2000). Transformation was performed using the Frozen-EZ Yeast Transformation kit (Zymo research). For excision assays, transformants were grown in 5 ml liquid media of CSM -leu-ura with 2% dextrose. After growth to saturation (36 hours), cells were washed twice with 5 ml water, resuspended in 0.5 ml water and plated onto CSM -his-leu-ade with 2% galactose. Colonies were counted after incubation at 30°C for 15 days and viable counts were made by plating 100 µl of a 1×10^5 and 1×10^6 dilution on YPD plates.

For the reintegration assay, cells were grown to saturation in 5ml liquid CSM -leu-ura with 2% dextrose, cells were washed twice with 5 ml sterile water, resuspended in 0.5 ml water and plated onto CSM -leu-ura-ade with 2% galactose plate and CSM -his-leu+5-FOA with 2% galactose plates. Colonies were counted after incubation at

30°C for 15 days, and viable counts were made by plating 100 µl of a 1×10^5 and 1×10^6 dilution on YPD plates. In another approach, individual Ade$^+$ *Muta1HIS* excision revertant colonies isolated directly from plates of CSM -his-leu-ade with 2% galactose were streaked on CSM -his+5-FOA plates to calculate the reintegration frequency.

Excision and reinsertion analysis

For footprint analysis, colony PCR was performed on *ADE2* revertant colonies using primers (Table S 1-3) flanking the insertion sites. PCR products were gel extracted (Zymoclean Gel DNA Recovery Kit) and sequenced. For reinsertion analysis, transposon display was conducted (Biedler et al., 2003). Genomic DNA was extracted from revertant colonies using the Yeastar genomic DNA kit (Zymo research); DNA samples were digested by *Bfa*I followed by adapter ligation. Pre-amplification and selective amplification were used to amplify the sequences between *Muta1* TIR and the *Bfa*I adapter sequence. Amplicons consisting of flanking sequences of the reinsertion sites and part of *Muta1* TIR were resolved on a 4% agarose gel, and polymorphic fragments were recovered and sequenced. Flanking sequences were mapped to the yeast genome (S288C, http://yeastgenome.org/) and the reinsertion sites were determined with regard to the closest genomic features. The insertion site analysis figure was made using the program Pictogram http://genes.mit.edu/pictogram.html.

Mutagenesis of *Muta1* transposase

Site-directed mutagenesis was used to generate mutant versions of *Muta1* transposase. One pair of primers was used for each mutation site, and pGAL415-ccdbMuta1 plasmid was used as template. PCR products were digested with *Dpn1* to remove template, and the resulting plasmid was sequenced to confirm that mutations occurred as expected.

Figures and Tables

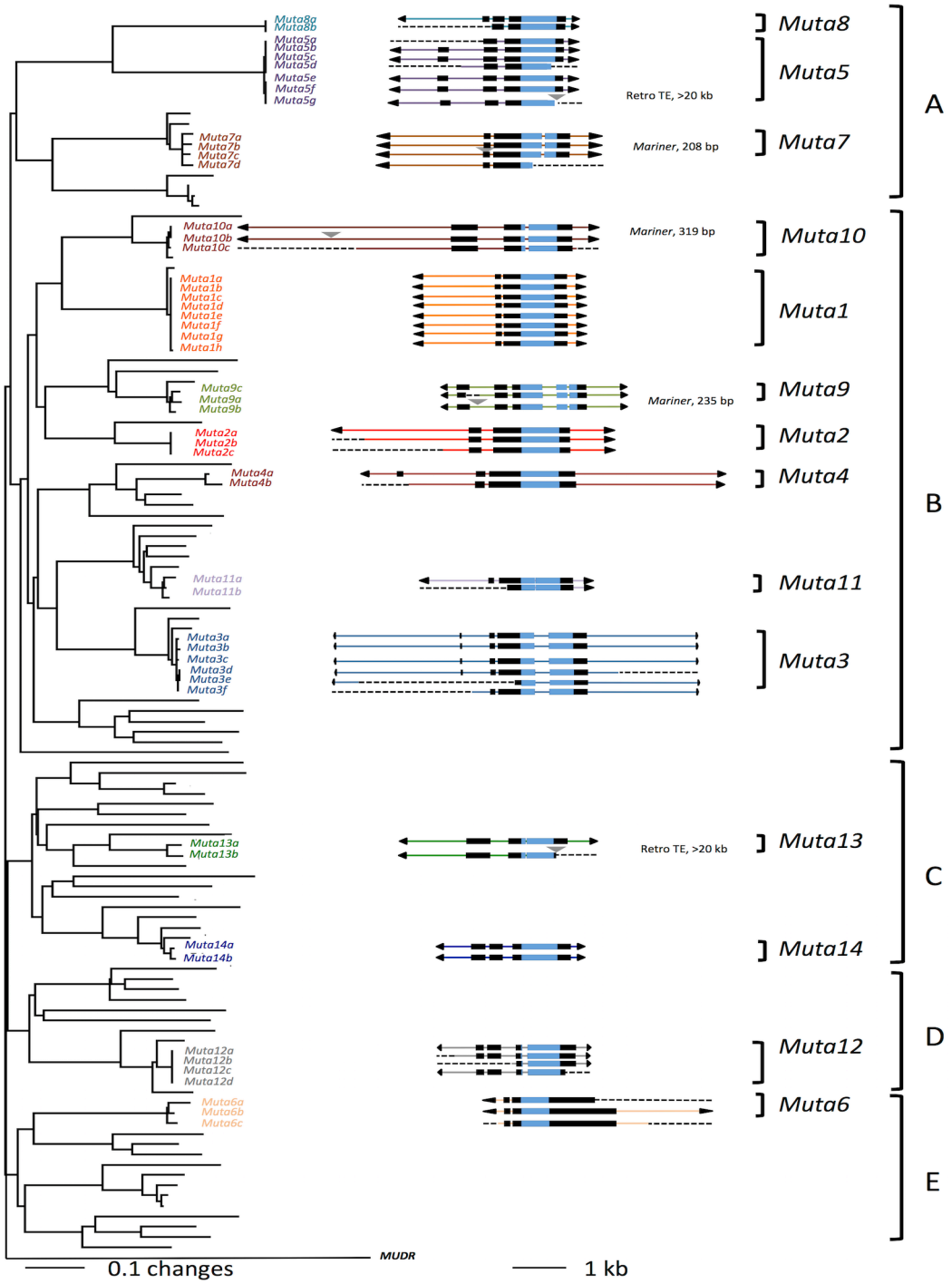

Figure 1-1. Phylogentic tree and structure of MULEs in *A. aegypti.*

Neighbor-joining tree generated from a multiple alignment of 141 conceptually translated catalytic domains from transposase proteins with bootstrap values calculated from 1000 replicates. For element structures: TIRs are black triangles, exons are black boxes, DDE domains are blue boxes, introns are lines connecting boxes, colored lines indicate within family identity of noncoding regions, other TE insertions are gray triangles above elements, and dashed lines are missing sequences caused by gaps, deletions or large insertions. The maize *MuDR* transposase is used as an outgroup. The *A. aegypti Mutator* transposases are classified into 5 major lineages (A-E).

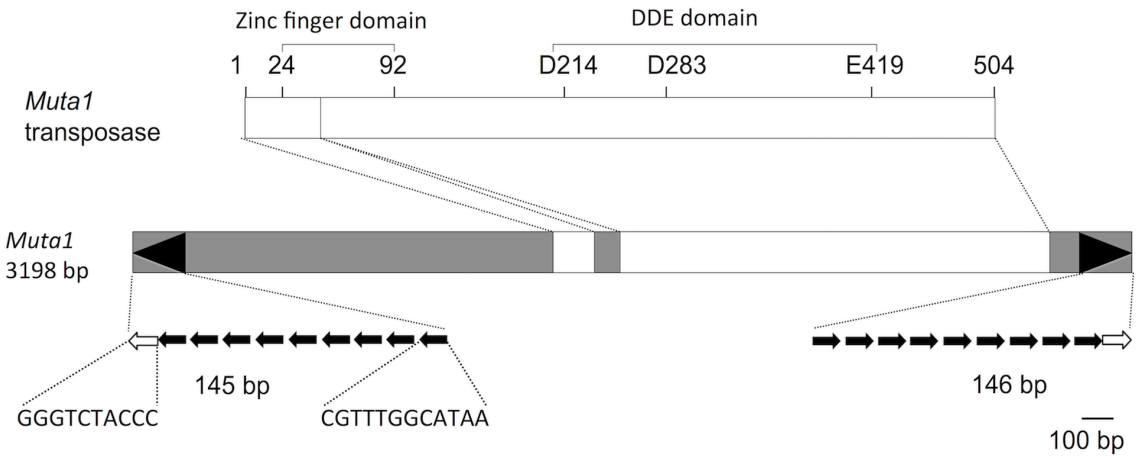

Figure 1-2. Structural features of the *Muta1* element and its transposase.

The eight virtually identical *Muta1* elements contain noncoding regions (shaded) and the coding region (white boxes) for the predicted 504-residue transposase with the predicted zinc and catalytic (DDE) domains discussed in the text. Structural features of *Muta1* include its distinctive long TIR (black arrowheads) whose substructure, expanded at the bottom, includes the 10 bp terminal palindromic motif (open arrow) and the 12 bp subterminal tandem repeats (black arrows) with linker DNA of 3-4bp represented by gaps between solid arrows.

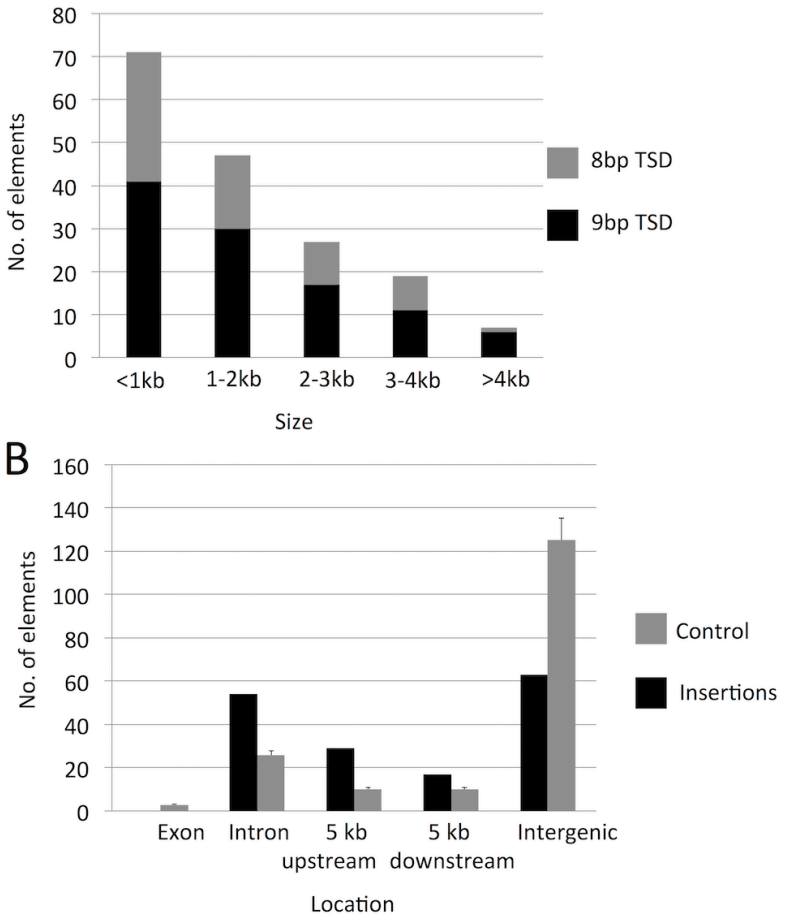

Figure 1-3. Features of *Muta1* derivative elements in *A. aegypti*.

(A) 171 *Muta1* derivative elements divide into 5 groups based on size. Within each group, the number of elements with 8 bp TSD and 9 bp TSD are shown in grey and black, respectively.

(B) Distribution of *Muta1* derivative element insertion sites in the *A. aegypti* genome. Mean ± s.d., n=1,000 (for control). Number of Insertion sites in the *A. aegypti* genome and control data set is shown in black and grey, respectively.

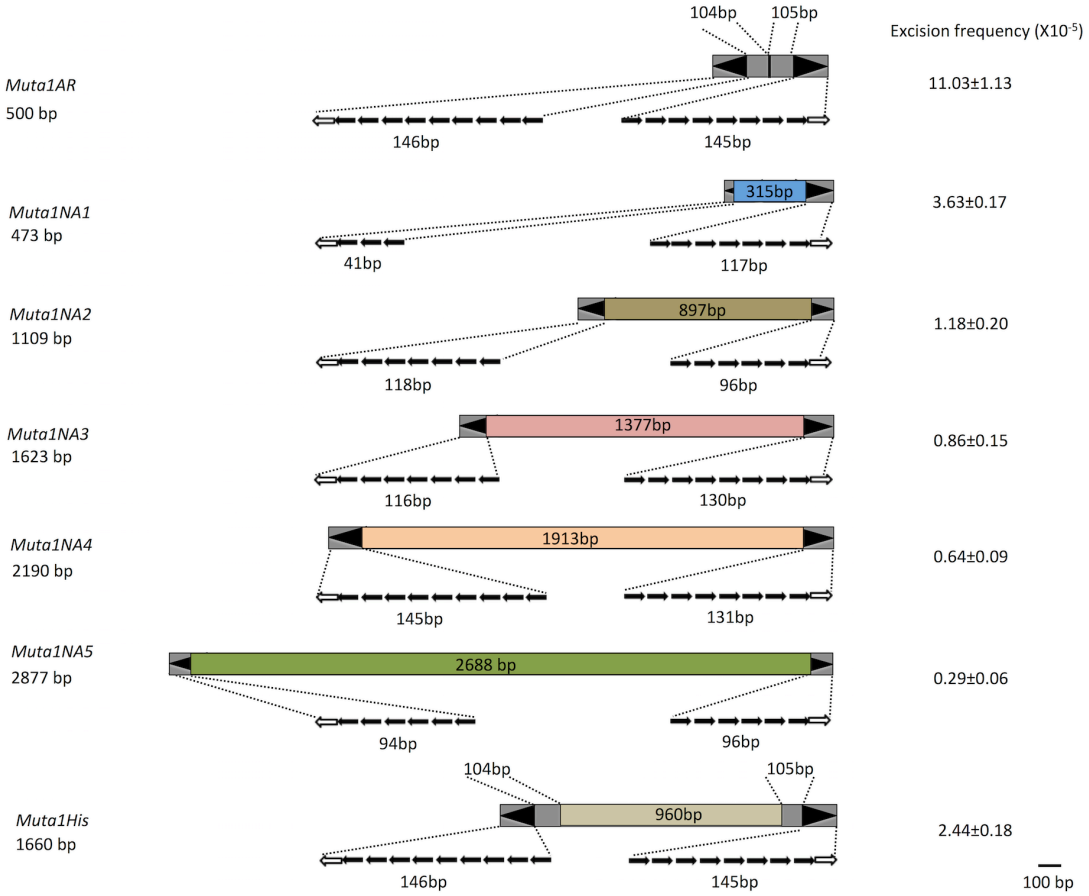

Figure 1-4. Structural features of nonautonomous *Muta1* elements used in this study.

Muta1NA1 through *Muta1NA5* are natural elements; *Muta1AR* and *Muta1His* are artificial. Element lengths and internal sequences are indicated. Black arrowheads represent the TIRs, which include the terminal palindromic motif (open arrow) and subterminal tandem repeats (solid arrows). Gray shaded regions are sequences derived from *Muta1*; colored regions of each element indicate the diverse origin of internal sequences. Excision frequencies from the *ADE2* reporter in yeast assays are on the right.

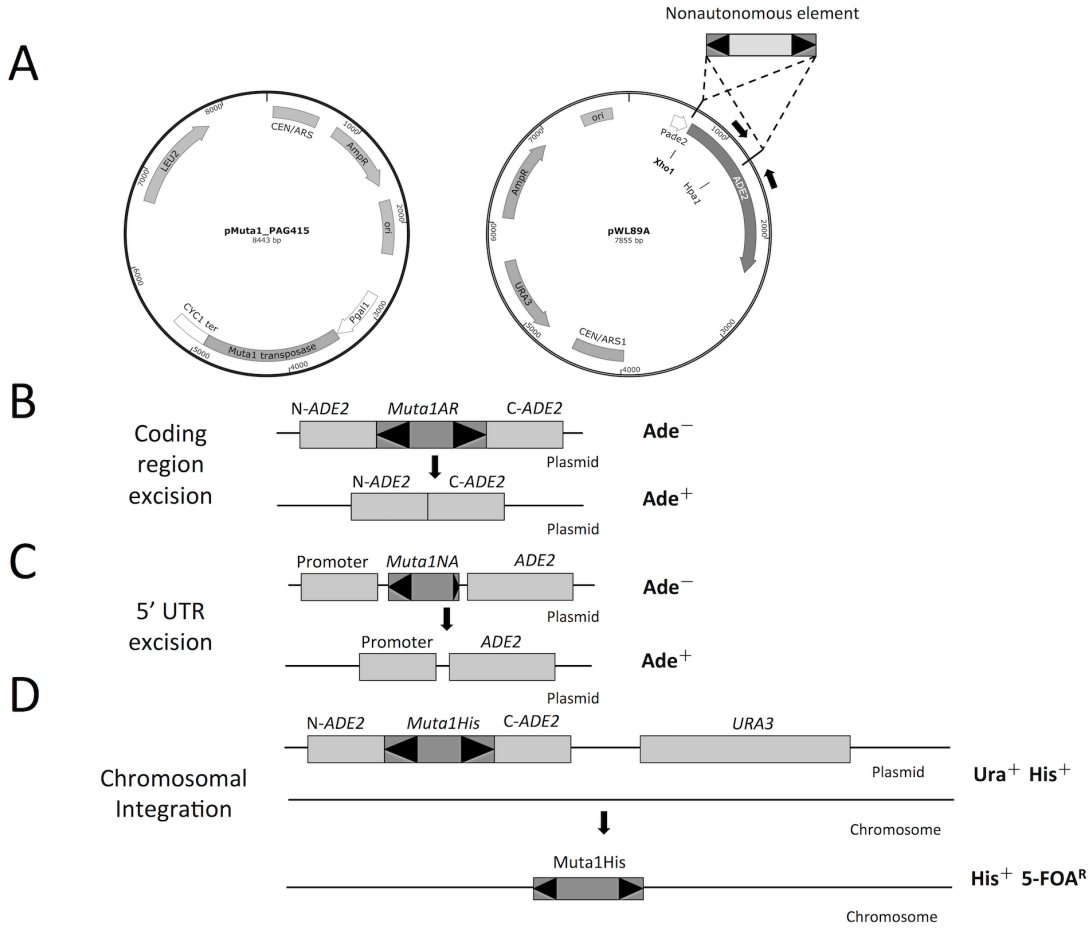

Figure 1-5. Yeast transposition assay constructs.

(A) The structure of pMuta1_PAG415 and pWL89A. AmpR, ampicillin resistance gene; ori, *E. coli* replication origin; Pgal1, GAL1 promoter; CYC1 ter, terminator; CEN, centromere sequences of yeast chromosomes; ARS, autonomous replication site. Dashed lines indicate the position of nonautonomous element insertions, in the 5'UTR and coding region respectively. Black arrows indicate the positions of primers used for PCR analysis in Figure S3A.

(B) Excision from coding region of *ADE2*.

(C) Excision from 5' UTR of *ADE2*.

(D) Reintegration. In the parental strain, pWL89A carries *Muta1HIS* in the coding region of *ADE2*. Reintegration is assayed by selecting cells that retain the *HIS* marker in Muta1HIS when the parental plasmid is excluded by 5-FOA treatment, which is toxic to Ura$^+$ cells.

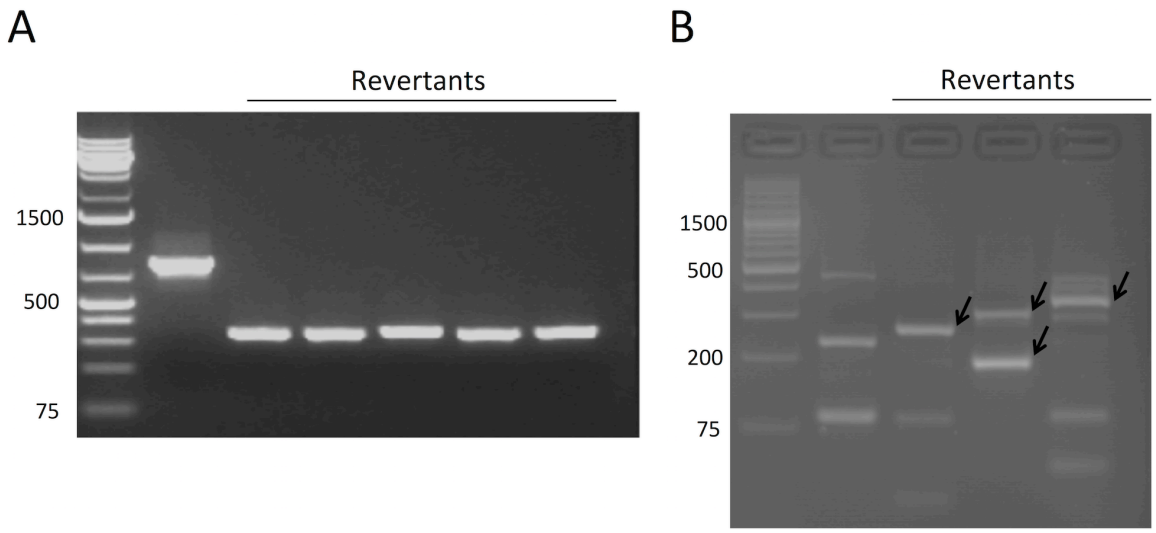

Figure 1-6. Analysis of excision and reinsertion events.

(A) PCR analysis of the *Muta1NA1* excision sites from ADE2 revertants using flanking primers. Expected band size is 820 (control) or 350bp, with or without *Muta1NA1,* respectively.

(B) Transposon display analysis of *Muta1HIS* reinsertion in yeast genome. DNA bands are amplicons consisting of flanking sequences of the reinsertion sites and part of TIR. PWL89A-Muta1HIS vector is used as control. Arrows indicate the polymorphic bands that represent the insertion of *Muta1HIS* in different genomic locations.

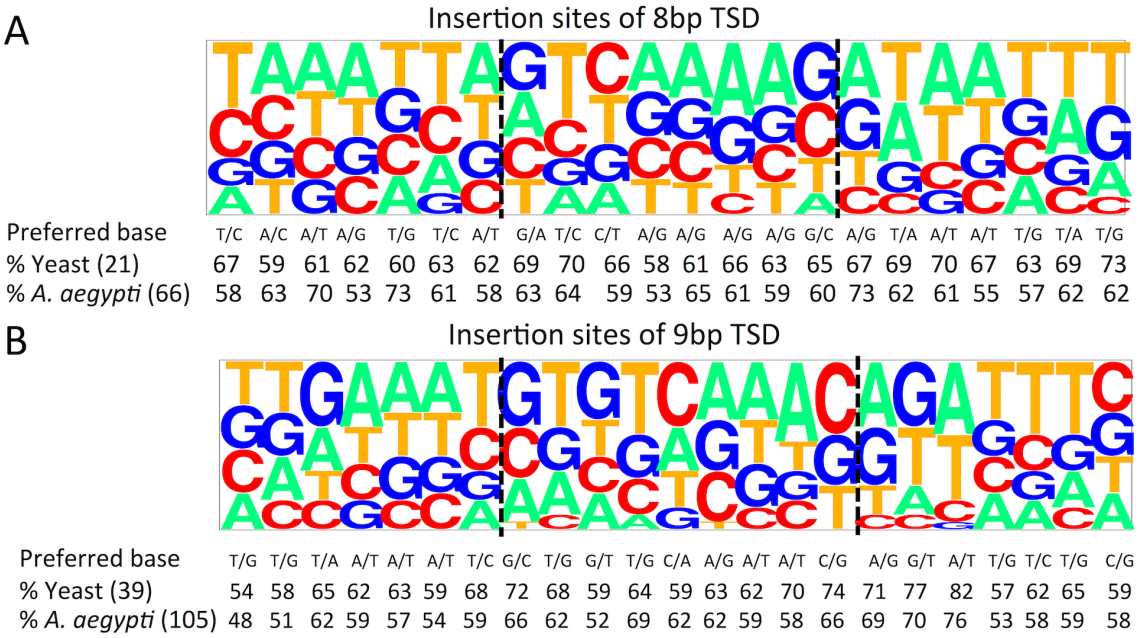

Figure 1-7. Seqlogo of insertion sites of *A. aegypti Muta1* derivative elements and reintegration sites in yeast.

Both 8bp TSDs (A) and 9bp TSDs (B) and their 7bp flanking sequences are analyzed, insertion preference is shown as a pictogram (height of letter indicates percentage of each nucleotide at that position) and the frequencies of preferred nucleotides, if any, are shown.

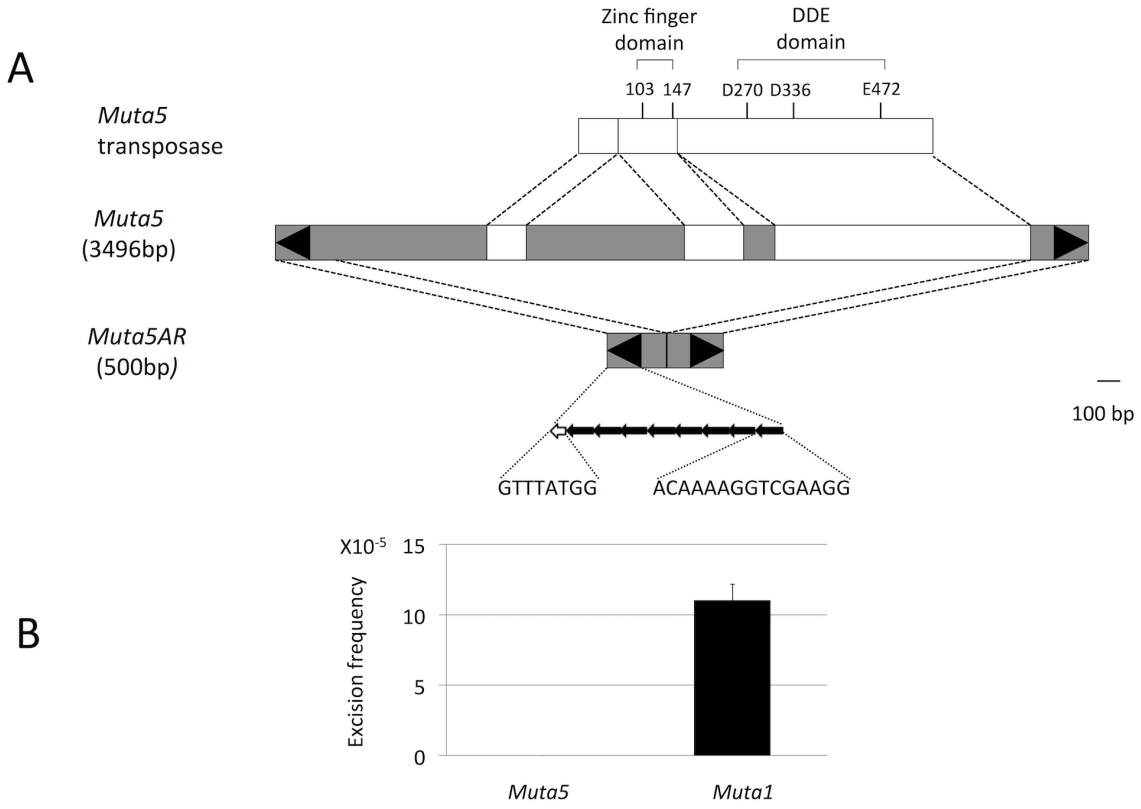

Figure 1-8. Structural features of *Muta5* and excision assay results.

(A) Structural features of *Muta5*. White boxes are coding regions, shaded boxes are noncoding regions, and triangles are the TIR. Within the TIR, black arrows represent 9 copies of the 15bp subterminal tandem repeat, open arrow represents the 8bp terminal motif. The putative 554-residue transposase is predicted to harbor a zinc finger domain and the catalytic (DDE) domain. The artificial *Muta5AR* element contains 250 bp from each end of *Muta5*.

(B) Excision frequencies of *Muta5AR* and *Muta5AR* from the *ADE2* reporter in yeast assay. *Muta1* is used as the positive control.

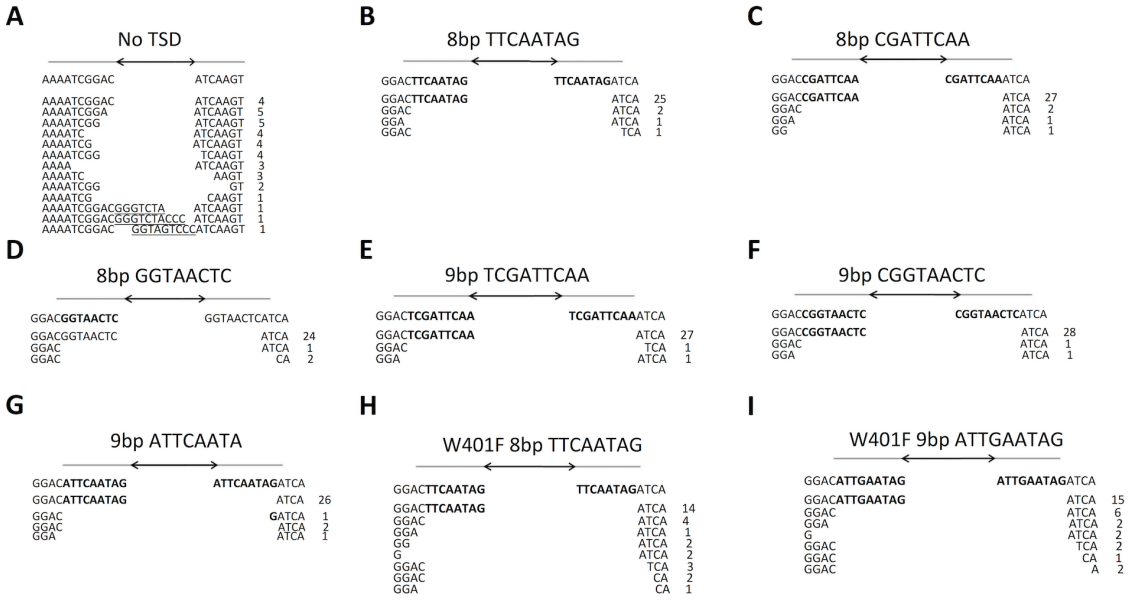

Figure 1-9. Footprints from *Muta1NA1* excision events.

Arrows indicate the *Muta1NA1* insertion, length and sequence of TSD in each assay is shown above the arrows. TSD or sequences derived from TSD are shown in bold, sequences derived from *Muta1NA1* are underlined, the number of recovered events are on right.

(A) Footprints of *Muta1NA1* excision from the *ADE2* 5' UTR without donor site TSD.

(B-D) Footprints of *Muta1NA1* excision from the *ADE2* 5' UTR with different 8 bp TSD sequence. (B): TTCAATAG; (C): CGATTCAA; (D): GGTAACTC.

(E-G) Footprints of *Muta1NA1* excision from the *ADE2* 5' UTR with different 9 bp TSD sequence. (E): TCGATTCAA, (F): CGGTAACTC, (G): ATTCAATAG.

(H-I) Footprints of *Muta1NA1* excision from the *ADE2* 5' UTR with the transposase W401F mutation. (H): the 8bp TSD TTCAATAG was used. (I): the 9bp TSD ATTGAATAG was used.

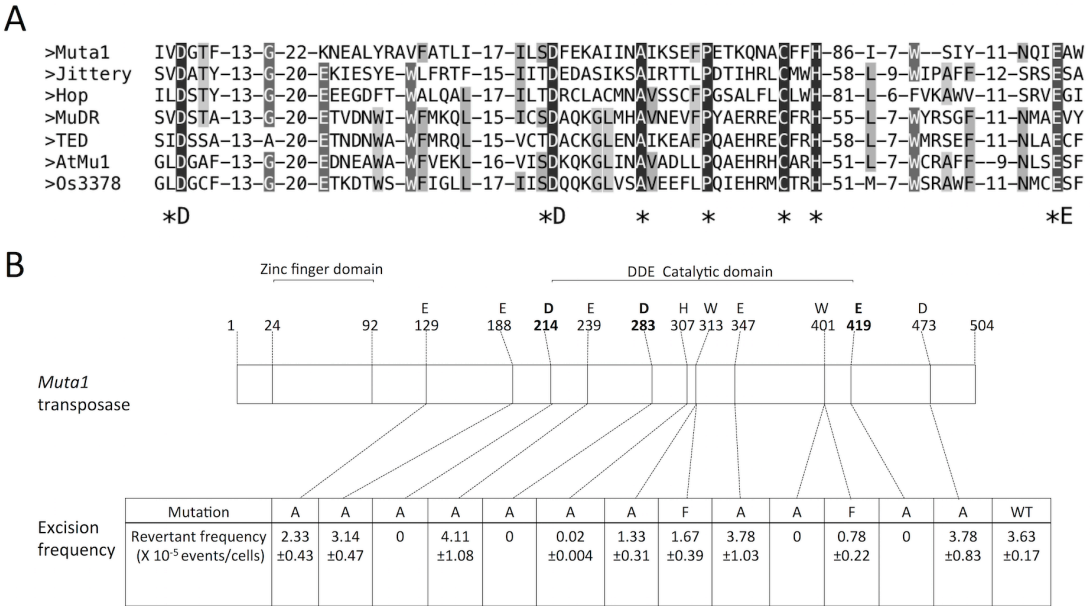

Figure 1-10. MUSCLE alignment of the DDE domain in MULE transposases and the impact of mutations.

(A) MUSCLE alignment of the DDE domain in *Muta1, MuDR, TED, atMu1, Os3378, Jittery, and Hop*. Shaded residues have related physical or chemical properties with darker shading denoting more conservation. Asterisks denote residues conserved in all sequences.

(B) Schematic of the 504 amino acid Muta1 transposase with positions of the putative FLYWCH type zinc finger DNA binding domain (24-92 aa), and DDE triad (residues D214, D283, E419). Designated amino acids were mutated to alanine (A) or to both alanine and phenylalanine (F) resulting in the excision frequencies shown below. See text for details.

Table 1-1. *Muta1* integration in yeast.

Assay	Frequency
Muta1His integration *	
ade2$^+$ excisant	2.44×10^{-5} Ade$^+$ cells/total cells
His$^+$ 5-FOA^R excisant	2.71×10^{-6} His$^+$ 5FOA^R cells/total cells
	ratio of reintegration = 11%
Muta1His integration **	
ade2$^+$ excisant	300 colonies
His$^+$ 5-FOA^R excisant	41 colonies
	ratio of reintegration = 14%

Table 1-2. Impact of TSD on transposition.

TSD	TSD sequence	Excision frequency (X 10^{-5})	Reintegration frequency (%) *	precise /total excision
0bp		11.03±1.13	13.9±1.8	4/38
8bp	TTCAATAG	63.34±9.12	20.88±1.53	25/29
	CGATTCAA	76.33±11.53	18.22±2.90	27/31
	GGTAACTC	68.53±12.16	19.15±2.39	24/27
9bp	ATTCAATAG	31.52±5.27	27.83±3.01	27/29
	TCGATTCAA	36.43±6.12	24.35±1.98	28/30
	CGGTAACTC	28.38±3.11	29.01±2.76	26/30

* Measured by analysis of independent Ade^+ revertants to maintain *Muta1His* but lacking plasmid.

Table 1-3. Comparison of *Muta1* and *Os3378* transposition in yeast.

	Muta1	*OS3378*
Excision frequency no TSD (X 10^{-7}) *	1103	1.2
Excision frequency 8bp TSD (X 10^{-7}) *	6940	–
Excision frequency 9bp TSD (X 10^{-7}) *	3211	20.2
Reintegration frequency No TSD **	13.90%	59.26%
Reintegration frequency 8bp TSD **	19.41%	–
Reintegration frequency 9bp TSD **	27.06%	39.28%
Percentage of reinsertion in gene rich regions	80%	68%

* Based on the excision of *Muta1AR* or *Os3378NA469* from the coding region of *ADE2*.

** Based on the reintegration of *Muta1His* or *Os3378NA469*.

– Corresponding experiment was not reported.

Chapter 2 Transposition of *Mutator*–like transposable elements (MULEs) resembles *hAT* elements and V(D)J recombination system

Abstract

Mutator–like transposable elements (MULEs) are widespread across fungi, plants and animals. Despite their abundance and importance as genetic tools in plants, the transposition mechanism of the MULE superfamily was previously unknown. The active *Muta1* element permits *in vitro* analysis of key steps in *Muta1* transposition using purified transposase. *Muta1* transposase specifically binds to the *Muta1* ends and catalyzes the excision of *Muta1* ends from donor site flanking DNA through double strand breaks (DSB), and the joining of newly excised transposon ends with target DNA. In the process, the DSB forms hairpin intermediates on the flanking DNA side. Analysis of mutant transposase revealed the involvement of the conserved DDE motif and a W residue. The transposition pathway resembles the *hAT* transposons and the V(D)J recombination system. The importance of a conserved W in transposition also supports the close relationship between MULE and *hAT* superfamilies. Yeast transposition and *in vitro* assays demonstrated that the terminal motif and subterminal repeats of *Muta1* TIR also influence *Muta1* transposition.

Introduction

Transposable elements (TEs) are DNA sequences that can move from one locus to another in the genome. TE induced genomic alterations, including insertions, deletions, duplications and translocations have been linked to changes in gene structure and expression (Finnegan, 1989, Kazazian et al, 2004). TEs are found in virtually all eukaryotes where they can comprise the largest proportion of the genome. For example, sequences derived from TEs account for about half of the human genome and more than 70% of the genomes of some grass species (Feschotte et al., 2007).

TEs are classified by their transposition intermediate: class 1 TEs use RNA whereas the element itself is the transposition intermediate for class 2 TEs. Class 2 TEs are classified into superfamilies on the basis of the element encoded transposase that catalyzes their movement (Wicker et al., 2007). A TE family consists of transposase-encoding autonomous elements and nonautonomous elements that do not encode functional transposase (Feschotte et al., 2007). Within a family, elements usually share the terminal inverted repeat (TIR) sequence and the length of the target site duplication (TSD) that is generated upon insertion by the action of transposase.

Mutator–like transposable elements (MULEs) are class 2 TEs first identified in maize (Robertson 1978; Lisch 2013) and subsequently found to be widespread with

members in plants, animals, protozoans, and fungi (Neuveglise et al., 2005; Pritham et at., 2005; Lopes et al., 2009; Marquez et al., 2010). Typical structural features of MULEs include long TIRs (>100bp) and 8-10bp flanking TSDs (Lisch 2002). Nonautonomous MULEs often carry various sequences between their TIRs including host gene fragments; such elements are called Pack-MULEs (Jiang et al., 2004). The high transposition frequency of one superfamily member, *MuDR* has been widely exploited for gene tagging and mutagenesis in maize (Lisch et al., 2009).

MULE transposases harbor a zinc finger DNA binding domain and a DDE catalytic domain (Babu et al., 2006; Nesmelova et al., 2010). Phylogenetic analysis of transposases from all known eukaryotic superfamilies indicate that MULEs and *hATs* are closely related. For example they share a [C/D](2) H motif located 15-45 amino acids downstream of the second D of the DDE triad and a W residue upstream of the E residue of the DDE triad (Yuan et al., 2011; Chalvet et al., 2004; Li et al., 2013). MULE and *hAT* elements also have longer TSDs than most other superfamilies, with the predominant TSDs for hAT elements at 8 bp, and 8 to 10bp for MULE elements. These shared features suggest that their transposition mechanism may also be similar.

Among the eukaryotic class 2 TEs, transposition begins with the cleavage of one strand at both ends of the element, the nucleophile is usually H_2O and this first cleavage results in an exposed 3′-hydroxyl groups (OH) at either the element side or

the flanking (host DNA) side (Figure 2-1). The free 3′OH serve as nucleophile for the cleavage of the second strand and liberation of the transposon from the donor site (Nesmelova et al., 2010; Hickman et al., 2010). For the second strand cleavage, there are three distinct mechanisms known to date. In the first pathway, Tc1/*mariner* transposases make double-stranded cuts to release the transposon end from donor site (Figure 2-1A, Plasterk et al., 1999; Feng et al., 2007). In the second pathway, first strand cleavage results in a 3′OH on the flanking DNA, which subsequently attacks the opposite strand to generate a hairpin structure on the flanking DNA (Figure 2-1B). This pathway is used by *hAT* and *Transib* elements, and the V(D)J recombination system, which produces the highly diverse repertoire of T cell receptors, antibodies and immunoglobulins (Hiom et al., 1998; Zhou et al., 2004, Henken et al., 2012). In the third pathway employed by *piggyBac* elements, the 3′OH generated in the first strand cleavage is on the transposon end. When it attacks the opposite strand, a hairpin structure is formed at the transposon end (Figure 2-1C, Mitra etal., 2008). After release from the donor site, the 3' OH on the free transposon ends serve as nucleophiles in a concerted attack on the target DNA in a reaction called strand transfer. The break in the target DNA is repaired by host enzymes, thereby generating the TSD on both sides of the inserted transposon. In contrast, the break in the donor site is repaired either by end-joining, a pathway that often leaves "footprints" at the donor site reflecting element excision (Weil et al., 2000), or homology-dependent gene conversion using a sister chromatid or homolog as a template (Engels et al., 1990; Plasterk et al., 1991).

Because MULEs and *hATs* have a close evolutionary relationship and share features including the[C/D](2)H motif, the W residue and long TSDs, their transposition mechanism is likely to be similar. To test this hypothesis, we studied transposition of *Muta1*, isolated from the genome of the mosquito, *Aedes aegypti,* which can spread dengue fever, yellow fever, chikungunya, zika, and many other diseases (Womack 1993; Marchette et al., 1969). *Muta1* transposase was able to catalyze the transposition of natural and artificial nonautonomous elements in a yeast assay (see chapter 1). In this study, steps of the *Muta1* transposition mechanism were dissected, from the binding of purified transposase to the *Muta1* end, to cleavage of the *Muta1* end from the donor site, and finally the joining of cleaved *Muta1* end to target DNA. Mutagenesis analysis revealed that the conserved DDE triad and the W residue shared with *hATs* play critical roles in all steps of *Muta1* transposition. The terminal motif and subterminal repeats of the *Muta1* TIR also impact transposition reactions.

Results

Muta1 transposase binds the transposon end

The *Muta1* transposase harbors a N-terminal FLYWCH type zinc-finger domain and a C-terminal DDE catalytic domain (Figure 2-2A). The 504 amino acid transposase was cloned, expressed and purified from *E. coli* as a His-tagged derivative. Initially, C-terminal His-tag was used, however the fusion protein had poor affinity for the nickel column. Therefore a N-terminal His-tag was added and the double His-tag fusion protein was used in subsequent experiments (see Methods).

Muta1 is 3198bp, flanked by 8 bp or 9 bp TSDs, and 145 bp TIRs comprised of a 10 bp imperfect palindromic terminal motif and 9 copies of a 12 bp subterminal tandem repeat separated by 3-4 bp linkers (Figure 2-2A). The *Muta1* L-TIR and R-TIR share high sequence identity (97.2%), with all sequence differences located in the linker DNA between subterminal repeats. For this reason the L- and R-TIR are considered as symmetric and only L-TIR was used in this study.

Before determining transposition reaction intermediates (Figure 2-1), we tested whether *Muta1* transposase (purified from *E.coli)* could bind to the L-TIR. A 5' P^{32} end radiolabeled 165 bp fragment containing the full length L-TIR and 20 bp of flanking sequence was used as substrate in DNA binding assays (Figure 2-2B). The result indicates that *Muta1* transposase binds to DNA fragments containing *Muta1*

L-TIR in the presence of 100 molar excess of nonspecific DNA. The binding is specific as addition of competitor DNA reduced the binding strength (Figure 2-2C, lanes 2-4). The many bands (ranging from 350 bp to greater than 700 bp) observed in *Muta1*-L-TIR binding possibly represent binding of oligomeric transposase proteins (Figure 2-2C, lanes 2-4).

Muta1 transposase catalyzes double strand breaks through a hairpin intermediate on the flanking DNA side

As shown in Figure 1, each of the characterized cleavage pathways produces distinct reaction intermediates that can be determined experimentally. For the donor cleavage assay, *Muta1* transposase was incubated with a 5' radiolabeled 188 bp end labeled fragment containing the full length *Muta1* L-TIR and 43 bp flanking sequence (Figure 2-3A). The reaction products were examined on both native and denaturing polyacrylamide gels to detect formation of a hairpin intermediate. On a native gel, two cleavage products were observed: the 145 bp band reflects the *Muta1* L-TIR and the 43 bp band reflects the flanking DNA (Figure 2-3B, lanes 5-8). The structure of these cleavage products is shown (Figure 2-3B diagram). Both fragments increase in amount as the incubation time is extended (from 10 to 120m, Figure 2-3B, lanes 5-8,).

On a denaturing gel, two reaction products were detected (Figure 2-3C, lanes 4-8). The size of the two products is consistent with the middle pathway in Figure 2-3A:

the 145 nt band reflects the *Muta1* L-TIR, and the 86 nt band reflects a linearized DNA hairpin containing both strands of the flanking DNA (43 + 43 nt, Figure 2-3C diagram and lanes 4-8). The amount of both fragments increases as incubation time is extended (from 5 to 120 min, Figure 2-3C, lanes 4-8,). These results indicate that *Muta1* donor cleavage involves formation of a hairpin structure in flanking DNA (Figure 2-3A, model 2).

Muta1 transposase joins the transposon end to target DNA

Donor site cleavage of *Muta1* generates free 3' OH on the L-TIR (Figure 2-3A, middle cleavage model), which could serve as a nucleophile in an end joining reaction. To test if *Muta1* transposase could join the *Muta1* L-TIR to target DNA, a "pre-cleaved", that is, with its 3' OH already exposed L-TIR fragment was constructed to serve as substrate with the target as intact pUC19 plasmid (2.6 kb). The substrate contains the full length L-TIR with 20 bp of *Muta1* internal sequence, on its bottom strand, the 3' end OH is exposed and the 5' end is radiolabeled (Figure 2-4A).

Labeled substrate and pUC19 plasmid were incubated with *Muta1* transposase and reaction products displayed on native and denaturing agarose gels (Figure 2-4). Two bands of 2.8 kb and 4.1 kb were observed on the native gel (Figure 2-4B, lanes 3-5). The 2.8 kb band reflects double-end join (DEJ), which is generated by concerted joining of L-TIR substrate to each strand of target DNA, thereby forming a linear, double-stranded DNA molecule (Figure 2-4B diagram). The 4.1 kb band

reflects single-end join (SEJ), which is generated by joining one L-TIR substrate to one strand of the target DNA. SEJ results in a nicked circular plasmid, in which one strand of pUC19 is broken by the joining of one L-TIR fragment and the other plasmid strand is intact (Figure 2-4B diagram). The amount of SEJ and DEJ product increases over time (from 20 to 120 min, Figure 2-4B, lane 3-5). On a denaturing gel, only one product is observed, consistent with DEJ and the linearized SEJ products, as the double-stranded SEJ product is denatured (Figure 2-4C, lane 3-5). These results indicate that *Muta1* transposase is able to join the *Muta1* L-TIR fragment to target DNA *in vitro*.

Subterminal repeats impact *Muta1* transposition reactions *in vitro*

After providing data for the mechanism of transposition, we turned our attention to the *cis* requirements for transposition beginning with the TIRs. *Muta1* can be classified as a type 2 foldback transposon because its TIR is composed of subterminal tandem repeats capable of forming secondary structures (Figure 2-2A, Engels et al., 1990). Although foldback elements are abundant in animal and plant species, in the MULE and *P* superfamilies, the function of the subterminal repeats is largely unknown (Plasterk et al., 1991; Windsor et al., 2000). With the active *Muta1*, we were, for the first time, able to experimentally test the function of subterminal repeats in foldback transposons. In the initial assays, DNA fragments containing the full length L-TIR (with 9 copies of the subterminal repeat) was used as substrate (Figure 2-4). The function of subterminal repeats was tested by reducing their copy

number in substrates used in DNA binding, donor site cleavage, and end joining assays.

DNA binding with full length L-TIR resulted in multiple bands, which suggest binding of oligomeric transposase (Figure 2-2C). To test if the binding of multiple transposase proteins correlated with the presence of multiple copies of subterminal repeats, we generated a series of DNA fragments consisting of variable copies of the subterminal repeat, and tested their binding activity. All six DNA substrates contain the same 20bp flanking sequence but different sequences derived from *Muta1* L-TIR, corresponding to the terminal motif alone; one copy of the subterminal repeat alone; the terminal motif plus 1, 2 or 3 copies of the subterminal repeat, and finally the full length TIR. Although the substrates vary in size (Figure 2-5A, diagram and bottom bands in lanes 1-12, 20-165 bp), their binding to transposase is specific as addition of competitor DNA reduced the binding strength (Figure 2-5A, lanes 1-2, 5-12). A single band around 350 bp was observed with the substrate containing the terminal motif and zero or one copy of subterminal repeat (Figure 2-5A, lanes 1&2, 5&6). No binding activity was detected with the substrate only containing the subterminal repeat and 20 bp flanking DNA (Figure 2-5A, lanes 3, 4). Additional bands (larger than 350 bp) were detected with substrates containing 2 or more copies of subterminal repeats (Figure 2-5A, lanes 7-12). These results indicate that the *Muta1* transposase binds specifically to fragments containing the 10bp terminal motif and 20bp flanking DNA (Figure 2-5A, lanes 1-2) but not to the fragment

containing only one copy of subterminal repeat and flanking DNA (Figure 2-5A, lanes 3-4). Increasing the number of the subterminal repeat promotes the binding strength and generates additional bands (Figure 5A, lanes 1-2, 5-12).

In the donor site cleavage assay, three different DNA fragments were used as substrates. They contain either or both of the terminal motif and the subterminal repeat, and 100 bp of flanking DNA (Figure 2-5B diagram). On a native polyacrylamide gel, the cleavage product (100 nt) was only detected with the substrate containing both motifs (Figure 2-5B, lane 3 and bottom diagram) after 60 min incubation with transposase.

In end joining assay, 3 different DNA fragments containing either or both of the 2 motifs and 20 bp of *Muta1* internal sequence were used as substrates (Figure 2-5C diagram). On native agarose gels, DEJ and SEJ products were only detected with the substrates containing the terminal motif, with or without the subterminal repeat. (Figure 2-5C, lane 2&6).

Terminal palindromic motif impacts *Muta1* transposition reactions *in vitro*

The sequence of the terminal motif of *Muta1* TIR, GGGTCTACCC, is an imperfect palindrome with the 5^{th} and 6^{th} nucleotides unpaired. To determine if the sequence and the palindromic pattern of this motif have any impact on transposition,

mutagenesis analysis was conducted for DNA binding, donor site cleavage and target joining reactions.

For the DNA binding assay, DNA fragments containing 20bp flanking sequence, one copy of the subterminal repeat, and mutant versions of the terminal motif were used as substrates (Figure 2-6A diagram) to achieve clear results and avoid the complication caused by the multiple nucleoprotein complexes (Figure 2-5A). Mutations introduced in the terminal motif were selected to disrupt or restore the palindromic pattern (Figure 2-6 A&B, red letters in sequences). For example, the first mutant has the first nucleotide G substituted by a T, and the second mutant DNA keeps this mutation and also has its 10th C changed to an A, thus restoring the palindrome. Band strengths indicate that mutations in the terminal motif have a negative impact on binding activity, however the weaker binding caused by a single mutation was partially restored by the second mutation, which restores the palindromic pattern (Figure 2-6A, lanes 3-12, Figure 2-6B). This indicates that the palindromic pattern is important for transposase binding. Another observation is that the weaker binding caused by single mutations is more apparent when the mutation located closer to the end of the TIR, for example, with mutation on 1st or 2nd nucleotide (Figure 2-6A, lane 3&5, Figure 2-6B), the binding is much weaker than the substrate with mutation on the 4th and 5th nucleotide (Figure 2-6A, lanes 9&11, Figure 2-6B).

In donor site cleavage assay, the DNA substrates (Figure 6C diagram) contain a 100bp flanking DNA segment, one copy of the subterminal repeat and the terminal motif with mutations (Figure 6C, red letters in sequences). Cleavage product (100 nt) was observed with all mutant substrates but less than with the wild-type terminal motif (Figure 2-6C, lanes 2-12), indicating that the mutations reduced cleavage activity, but the palindromic pattern had no significant impact.

In the end joining assay, substrates (Figure 2-6D diagram) contain one copy of the subterminal repeat and mutated versions of the terminal motif (Figure 2-6D, red letters in sequences). The amount of SEJ and DEJ products indicate that mutations in the 3rd and 4th/5th nucleotide (Figure 2-6D, lane 7&12) have negative impact on end joining activity, while mutations in other positions and the palindromic pattern has no significant impact (Figure 2-6D, lanes 2-12).

Subterminal repeats and terminal motif impact *Muta1* transposition in yeast

The subterminal repeats and the terminal motif impact *Muta1* transposition reactions *in vitro* (Figure 2-5&2-6). To determine if they also impact transposition *in vivo*, yeast transposition assays were performed. The assay employs a *Muta1* transposase expression vector and a reporter vector containing a nonautonomous element inserted in the 5' UTR that blocks expression of the *ADE2* gene (Figure 2-7A). Transposase-mediated element excision restores *ADE2* expression and permits the cells to grow on minimal plates (Figure 2-7B).

To test the contribution of subterminal repeats, we constructed a series of artificial nonautonomous elements and tested their mobility in yeast. The artificial elements were generated by modifying *Muta1NA1*, which is a natural *Muta1* derivative element cloned from the *A. aegypti* genome (Figure 2-8A). In the artificial elements, the 315bp internal sequence of *Muta1NA1* was flanked by variable numbers of the two motifs, as indicated by the number of "Ts" (Terminal motif) and "Rs" (subterminal Repeat). For example, 1T0R+1T7R represents an element with one copy the terminal motif but no subterminal repeat on the left end and one copy of the terminal motif and seven copies of the subterminal repeat on the right end. The transposition frequencies of the artificial elements indicate that having the terminal motif and one copy of the subterminal repeat on both ends is sufficient for an element to be mobilized by *Muta1* transpoase, and increasing the copy number of subterminal repeats promotes transposition (Figure 2-8B).

To test the impact of the sequence and palindromic pattern of the terminal motif, mutations were introduced at both ends of *Muta1NA1*. All mutations resulted in reduced excision frequency, and the palindromic pattern seems to be unimportant (Figure 2-8C). Notably, mutations closer to the end of the TIR led to a more severe reduction in transposition activity. For example, the excision frequency of *Muta1NA1* was lower when a mutation was located in the 1st or 2nd nucleotide than in the 4th and 5th nucleotide of the terminal motif, (Figure 2-8C).

Identification of the catalytic core of the *Muta1* transposase

Prior mutagenesis analysis identified several functionally important residues of *Muta1* transposase, including the DDE triad (D214, D283, E419) and W401 (see chapter 1), which is shared between the transposases from MULEs and *hAT* elements. To probe the function of these amino acids in *Muta1* transposition, transposases with D214A, D283A, E419A, W401A or W401F mutations were purified and assayed for DNA binding, donor site cleavage, and end joining activities.

Although still capable of binding specifically to the *Muta1* L-TIR fragment (Figure 2-9A, lanes 2-13), the alanine substitution mutant transposases lost activity in all DNA cleavage and joining reactions, while the W401F mutation retained partial activity. First, in the cleavage assay, no cleavage products were detected after 120 min incubation with alanine substitution mutant transposases (Figure 2-9B, lanes 3-6), and the W401F mutant generated less cleavage products than wild-type transposase (Figure 2-9B, lanes 2&7). The alanine substitution mutants were also defective in the end joining reaction (Figure 2-9C, lanes 3-6), as no joining products were detected, while the W401F mutation only lost partial activity (Figure 2-9C, lanes 2&7). These results indicate that the DDE triad and the W reside are involved in donor site cleavage and end joining of *Muta*1-mediated transposition.

Discussion

Our experiments provide evidence for a double strand break mechanism catalyzed by the *Muta1* transposase involving formation of a hairpin structure in the flanking DNA, thereby releasing the transposon from the donor site (Figure 1B&3C). The same mechanism has been reported for *hAT* and *Transib* elements, the V(D)J recombinase RAG1, and a subset of retroviral integrases (Figure 2-1B, Zhou et al., 2004; Hencken et al., 2012; Agrawal et al., 1998; Craig et al., 2002). These findings support the close evolutionary relationship between MULEs and *hATs,* first revealed by phylogenetic analysis (Yuan et al., 2011).

After transposon excision, the gap repair at the donor site requires the opening of the hairpin structure, our experiments suggests that this step is not likely performed by *Muta1* transposase, because the amount of the hairpin accumulates over time and the cleaved hairpin is not detected (Figure 2-3C). Similar finding is also observed in the *Hermes, Transib* and RAG1 systems (Zhou et al., 2004; Hencken et al., 2012; Agrawal et al., 1998). The hairpin may be opened by host enzymes like Artemis, which can open hairpins (Zhou et al., 2004; Agrawal et al., 1998; Mitra et al., 2008). Subsequent joining of the opened hairpins in *hAT, Transib* and RAG1 systems may be achieved by the nonhomologous end-joining pathway, which often leaves footprints at the donor site (Zhou et al., 2004; Agrawal et al., 1998; Weil et al., 2000). *Muta1* may utilize a distinct gap repair mechanism because yeast transposition assay revealed that the majority (> 90%) of excision events were

precise (see chapter 1), which means the element was removed as well as a single copy of the TSD, thus leaving no footprint at the donor site. Interestingly, variable *Muta1* footprints were detected when TSD was absent from the donor site (see chapter 1), similar to other MULE (rice *Os3378*) and the prokaryotic IS*256* family that is closely related to MULEs (Hennig et al., 2008; Zhao et al., 2015). It was hypothesized that IS*256* utilizes a transposase-independent replication slippage pathway for the donor site gap repair, which requires at least 8 bp of homologous DNA flanking the break point (Hennig et al., 2008). Therefore, the gap repair mechanism of *Muta1* is more likely similar to IS*256* family instead of *hAT* elements.

Mutagenesis analysis confirms the importance of the DDE triad in *Muta1* transposition reactions (Figure 2-9 A-C). Direct involvement of the DDE triad in DSB and end joining reactions has also been shown for *Hermes, Transib* and *piggyBac* transposases and for the RAG1 recombinase. In all cases, transposases with alanine substitution of the DDE triad lost activity for all reactions (Zhou et al., 2004; Agrawal et al., 1998; Hencken et al., 2012; Mitra et al., 2008). Prior studies showed that the conserved DDE triad of many transposases (including *hAT* Tc1/*Mariner, piggyBac* and the prokaryotic Tn5) and the RAG1 recombinase was organized in an RNaseH-like fold to form the active site. In the typical RNase H-like fold of β1- β2-β3-α1-β4-α2/3-β5-α4-α5/6, the DDE triad is usually located on specific secondary strucutres, with the first D on β1, the second D right after β4, and the E on α4 (Figure 2-10, RAG1, Hickman et al., 2010). However, the secondary structure

predictions suggest that the same fold is not likely present in most *A. aegypti* MULE transposases, because of the lack of β-strand adjacent to the second D of the DDE triad and the addition of 4 α-helices between the second D and the E (Figure 2-10). Possibly the multiple critical residues within *Muta1* transposase are closely positioned to form the active center through more complex folding or formation of transposase oligomer.

One explanation for the fact that MULE, *hAT, Transib* and V(D)J systems all form hairpin structures in flanking DNA is that besides the DDE triad, they share a few functionally important amino acids which are critical for the hairpin formation, especially the W residue. Crystallographic analysis and *in vitro* assays revealed that the W318 in *Hermes* plays an important role in the flanking DNA positioning, which ensures that DSB occurs at the correct position in element excision process (Hickman et al., 2014). And crystallographic analysis revealed that mutation on the corresponding W893 of RAG1 could destabilize the structure of RAG1 (Kim et al., 2015). We also identified the W from multiple alignment analysis (Figure 2-11). For *Muta1,* alanine substitution of W401A mutation abolished all DNA cleavage and joining reactions, while W401F caused reduction of activity (Figure 2-9 B&C) and decreased the ratio of precise excision from 90% to 50% in yeast assay (see chapter 1), indicating this mutation could cause inaccuracy in *Muta1* excision. Collectively, our data suggests this W plays a critical role in *Muta1* transposition, possibly involved in the hairpin formation, which is similar to its role in *hAT* elements.

Despite their abundance, the function of the subterminal repeats of Foldback transposons remains unclear (Rebatchouk et at., 1997; Potter et al., 1980; Windsor et al., 2000). It has been suggested the tandem repeats in a foldback transposon could enhance the terminal recognition of transposase by providing more binding sites (Marzo et al., 2013). However, this hypothesis has not been experimentally demonstrated because of the lack of active foldback elements. Because *Muta1* is active in heterologous system (yeast) and the transposase is capable of catalyzing transposition reactions *in vitro*, we had the first chance to test if the function of subterminal repeats in foldback transposons. In DNA binding assay, when the number of subterminal repeats increases in the substrates, the binding with transposase is not only stronger but also generates additional bands, this suggests that the multiple copies of subterminal repeats could enhance the binding with transposase (Figure 2-5A). In the donor site cleavage assay, cleavage product was only detected when the subterminal repeat is present in the substrate (Figure 2-5B). In addition, yeast excision assays revealed increased copy number of the subterminal repeats in the TIR had positive impact on excision frequency (Figure 2-8A). This could be explained by the stronger binding activity resulted from more copies of subterminal repeats. Collectively, our study provides the first experimental evidence that subterminal repeats are critical for transposition of foldback transposons, possibly by enhancing the transposase recognition.

The palindromic pattern of the *Muta1* terminal motif (sequence of GGGTCTACCC) only affects transposase binding (Figure 2-6A-D), some restriction enzymes also recognize palindromic sequences as their binding site (Sakai et al., 1995), *Muta1* termini recognition may utilize similar mechanism, in which the terminal palindromic motif could provide structural hints to delimit the end of the element. Mutations in the *Muta1* terminal motif also reduced donor site cleavage activity and excision frequency in yeast, suggesting its sequence may be optimal for transposition. Similarly, mutagenesis on the end of TIR also reduces transposition for other transposons, for example, when mutations were introduced to the TIR of maize *Ds* element, excision activity was greatly reduced (Weil et al., 2000).

Conclusions

We have dissected the mechanism underling *Muta1* transposition. Similar DNA breakage and target joining mechanism and conserved residues support the close evolutionary relationship between *MULE* and *hATs* superfamilies and the V(D)J recombination system. The involvement of subterminal repeats in foldback element transposition is also experimentally demonstrated. Future analysis of *Muta1* transposition may provide new insights to understand the evolutionary relationships between MULE, *hAT* elements and the V(D)J recombination system.

Materials and methods

Muta1 transposase expression and purification

The *Muta1* transposase coding sequence (1512 bp) was isolated by PCR amplification from pMuta1-PAG415 (Figure 2-7A, see chapter 1). In the PCR process, codons of 6 histidines were added to the N-terminal of the coding sequence and cloned between the *Nco*I and *Pst*I sites of plasmid pBAD-Myc-HisC (Invitrogen) (which has a C-terminal His-tag) to generate pBAD-Muta1-His-NC. This fusion construct has 6 His residues on both the N and C terminus of the *Muta1* coding sequence. *E. coli* Top10 (Invitrogen) cells containing pBAD-Muta1-His-NC were grown with shaking at 30°C in LB medium containing 100 mg/ml carbenicillin to OD600 of 0.5. The culture was induced with 0.2% L-arabinose for 18 h at 16°C. Cells were harvested by centrifugation at 8000 rpm for 5 minutes and resuspended in 8 mL of PBS buffer (50 mM NaH_2PO_4, 500 mM NaCl and 10 mM imidazole, pH 7.5). After addition of lysozyme (0.2g/liter of culture), cells were incubated on ice for 30 minutes before lysis by sonication for 30 min (Fisher Scientific). 6ml lysate was loaded onto a pre-equilibrated Ni^{2+} column (Invitrogen) containing 1ml Ni^{2+} resin. The column was washed with 35 ml of TBS buffer with 50 mM imidazole followed by 35 ml of TBS buffer with 100 mM imidazole. Muta1–His fusion protein was eluted with 7 ml of TBS buffer with 500 mM imidazole, dialyzed twice against 35 ml volume of TBS buffer and concentrated using centrifuge filtration (Amicon system) to a final volume of 100 ul. Glycerol was added to a final concentration of 15% (v/v) and stored at -80°C.

DNA binding assay

DNA fragments containing full length L-TIR were generated by PCR from full length *Muta1,* fragments corresponding to partial TIR were synthesized oligos (IDT). DNAs substrates were 5' end radiolabelled on both strands with gamma-P^{32}-ATP and T4 polynucleotide kinase (NEB). 120nM *Muta1* transposase and 2nM radiolabelled DNA were incubated in 25mM HEPES, 5% (v/v) glycerol, 0.01% bovine serum albumin (BSA) and 4mM dithiothreitol (DTT) in the presence of a 100-fold molar excess of sheared herring sperm DNA at room temperature for 30 minutes. Unlabeled full length *Muta1* L-TIR DNA was used in 100 or 500 molar excess as competitor DNA. The products were run on 5% native acrylamide gels, all gels were dried and exposed to X-ray film.

Double strand break reactions

The DSB assay was conducted as previously described. 200nM of *Muta1* transposase was incubated with 1.5nM of radiolabelled *Muta1* L-TIR or other end fragments in 25mM HEPES (pH 8.0), 3mM Tris (pH 8.0), 75mM NaCl, 2mM DTT, 10mM $MgCl_2$, 0.01% BSA, 5% glycerol and 10nM pUC19 in a final volume of 20 ul at 30°C for different time intervals. Reactions were stopped by incubation in 1% SDS and 20mM EDTA for 30 min at 65°C and displayed on 5% native and denaturing (with 7M urea) acrylamide gels.

End joining reactions

Following published protocols (Zhou et al., 2004; Mitra et al., 2008), 200nM of *Muta1* transposase was incubated with 1.5nM of radiolabelled *Muta1* L-TIR or other end fragments in 25mM HEPES (pH 8.0), 3mM Tris (pH 8.0), 75mM NaCl, 2mM DTT, 10mM MgCl2, 0.01% BSA, 5% glycerol and 10nM pUC19 in a final volume of 20 ul at 30°C for 30 minutes. Products were displayed on native or denaturing (with 50mM NaOH) 1% agarose gels.

Mutagenesis of *Muta1* transposase

Site-directed mutagenesis was used to generate mutant versions of *Muta1* transposase. One pair of primers was used for each mutation site, and PBAD-Muta1-His-NC plasmid (described in first section of Methods) was used as template. PCR products were digested with *Dpn1* to remove template, and the resulting plasmid was sequenced to confirm that mutations occurred as expected.

Yeast transposition assay

The yeast transposition assay using *Saccharomyces cerevisiae* strain DG2523 and the pWL89A vector were described previously (Yang et al., 2006; Weil et al., 2000). The pMuta1-PAG415 plasmid (Figure 2-7A, see chapter 1) was used for transposase expression. Mutant versions of *Muta1NA1* were generated by PCR. All nonautonomous elements were inserted in the *Xho*I site of pWL89A (Figure 2-7B) through homologous recombination in yeast as previously described (Yang et al.,

2006). Transformation was performed using the Frozen-EZ Yeast Transformation kit (Zymo research). Transformants were grown in 5 ml liquid media of CSM -leu-ura with 2% dextrose. After growth to saturation (36 hours), cells were washed twice with 5 ml H_2O, resuspended in 0.5 ml H_2O and plated onto CSM -his-leu-ade with 2% galactose. Colonies were counted after incubation at 30°C for 15 days and viable counts were made by plating 100 µl of a 1×10^5 and 1×10^6 dilution on YPD plates.

Colony PCR was performed on *ADE2* revertant colonies using primers flanking the insertion sites and PCR products were gel extracted (Zymoclean Gel DNA Recovery Kit) and sequenced for footprint analysis.

Figures

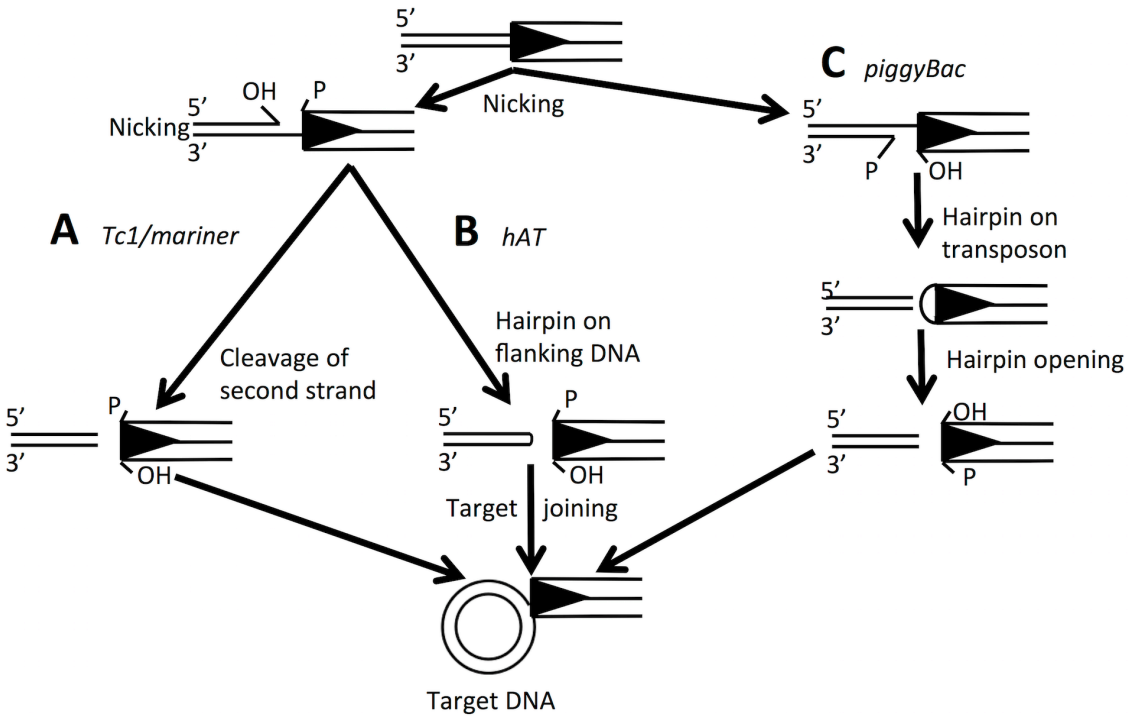

Figure 2-1. Comparison of transposase-mediated cleavage and target joining mechanisms.

Key steps in three DNA cleavage and target joining pathways of TE superfamilies. Black triangle indicates the TIR. After transposon ends released through different DNA cleavage mechanisms, the free 3' OH is used for target joining.

(A) Tc1/*mariner* transposases make double-stranded cuts and generate element end with free 3' OH.

(B) *hAT* transposases catalyze the cleavage reaction through formation of a hairpin structure in flanking DNA.

(C) *piggyBac* transposases generate a hairpin structure at the transposon end and release transposon from the donor site.

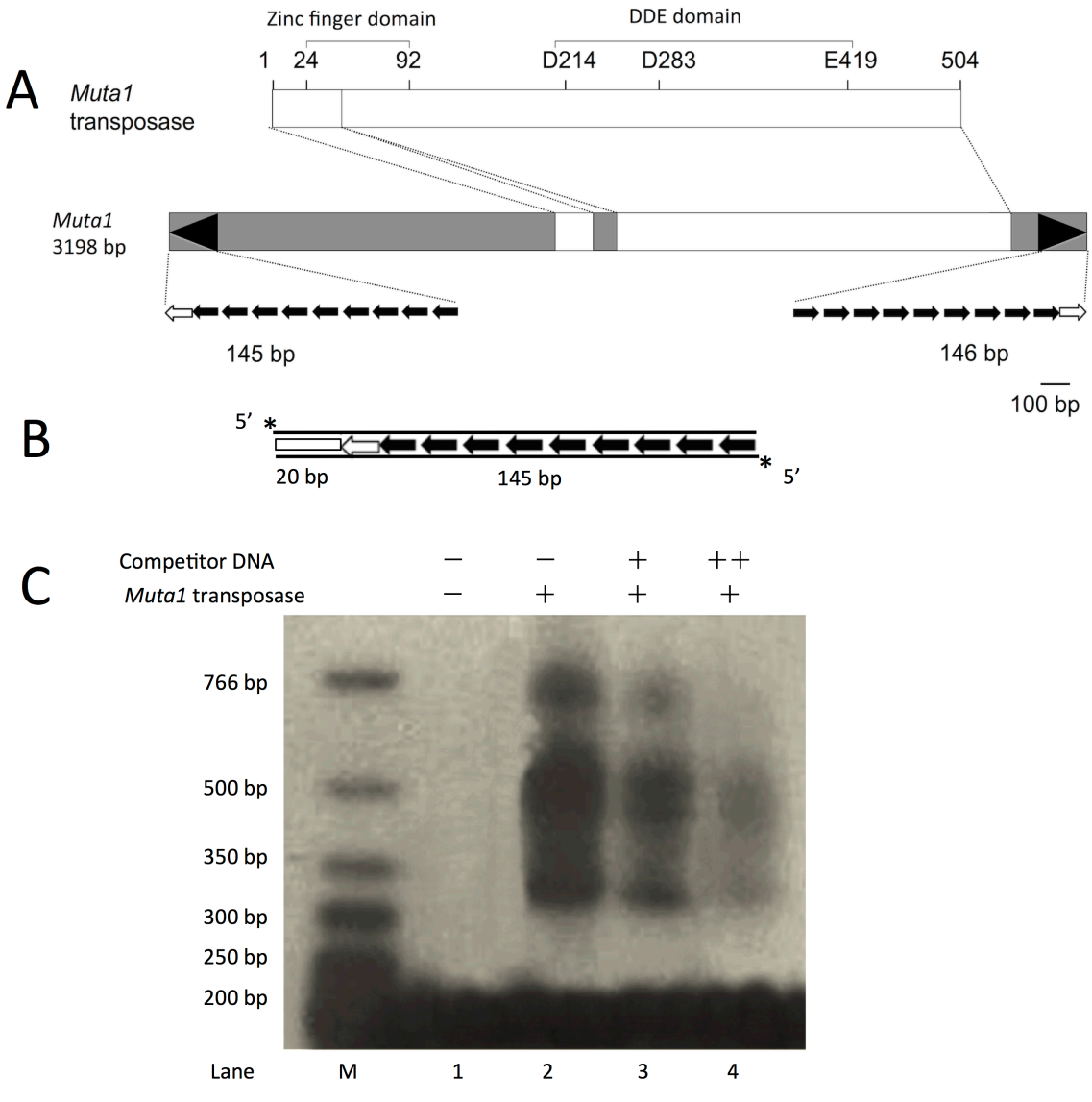
A
Muta1
transposase
Zinc finger domain
DDE domain
1
24
92
D214
D283
E419
504
Muta1
3198 bp
145 bp
146 bp
100 bp
B
5'
*
20 bp
145 bp
*
5'
C
Competitor DNA
Muta1 transposase
766 bp
500 bp
350 bp
300 bp
250 bp
200 bp
Lane
M
1
2
3
4

Figure 2-2. *In vitro* analysis of DNA binding of *Muta1* transposase.

(A) The 3198 bp *Muta1* encodes a 504 amino acid transposase with the predicted zinc and catalytic (DDE) domains. White boxes indicate exons and grey boxes indicate noncoding regions. Black arrowheads represent the TIR whose substructure is expanded below. In this and all subsequent figures: open arrows represent the 10 bp terminal palindromic motif, and black arrows are the 12 bp subterminal tandem repeat.

(B) The 165 bp DNA substrate used in DNA binding assay is 5' radiolabelled at both strands (asterisk) and contains a 20 bp flanking DNA segment (white box) and the full length *Muta1* L-TIR.

(C) *Muta1* tpase and DNA binding assay. Lanes 1-4: fastest migrating band present in all lanes is DNA substrate, slower migrating bands (> 300 bp) are nucleoprotein complexes. Competitor DNA is unlabeled DNA in (B) which was added in 100 (+) or 500 (++) molar excess. One hundred molar excess of nonspecific DNA is added in all reactions.

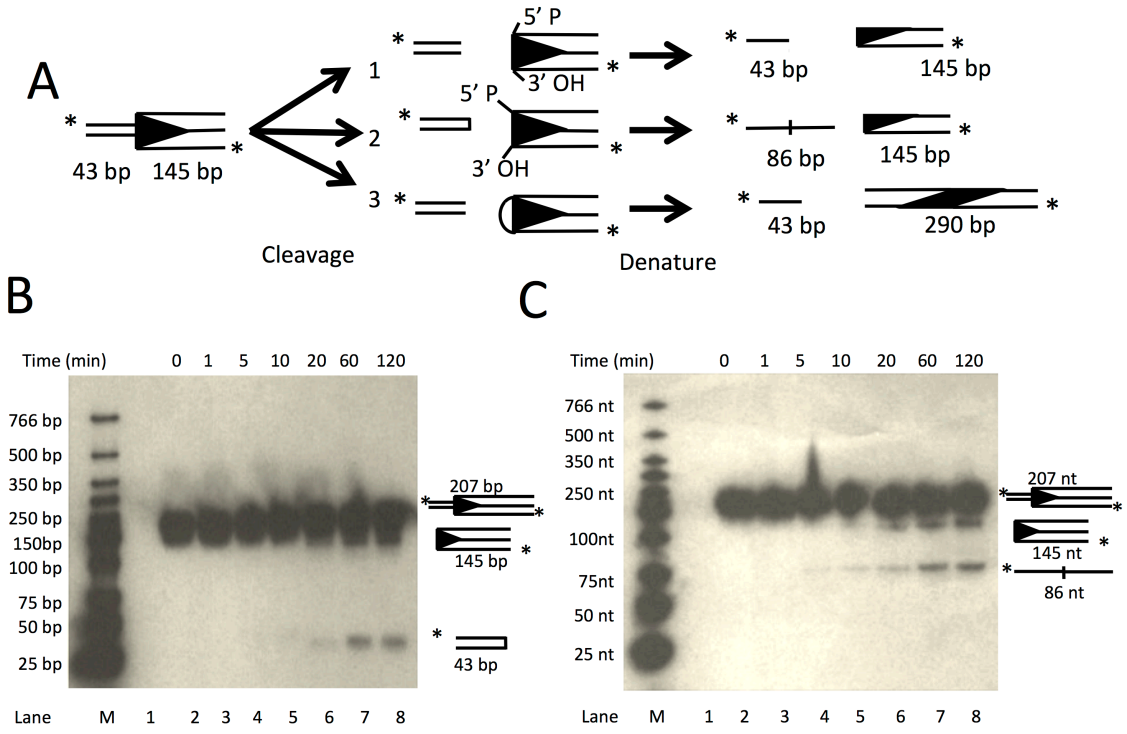

Figure 2-3. *In vitro* analysis of *Muta1* double strand cleavage.

(A) The 188 bp DNA substrate is 5' radiolabeled and contains full length *Muta1* L-TIR (145 bp) and a 43 bp flanking segment. The possible reaction outcomes of the three DNA cleavage pathways in Figure 1 are shown.

(B) Reaction products on a native polyacrylamide showing 43 bp and 145 bp *Muta1* L-TIR products (lanes 5-8).

(C) Reaction products on a denaturing polyacrylamide gel showing the 145 and 86 nt bands (lanes 4-8).Diagram on right indicates the predicted structure and size of each band.

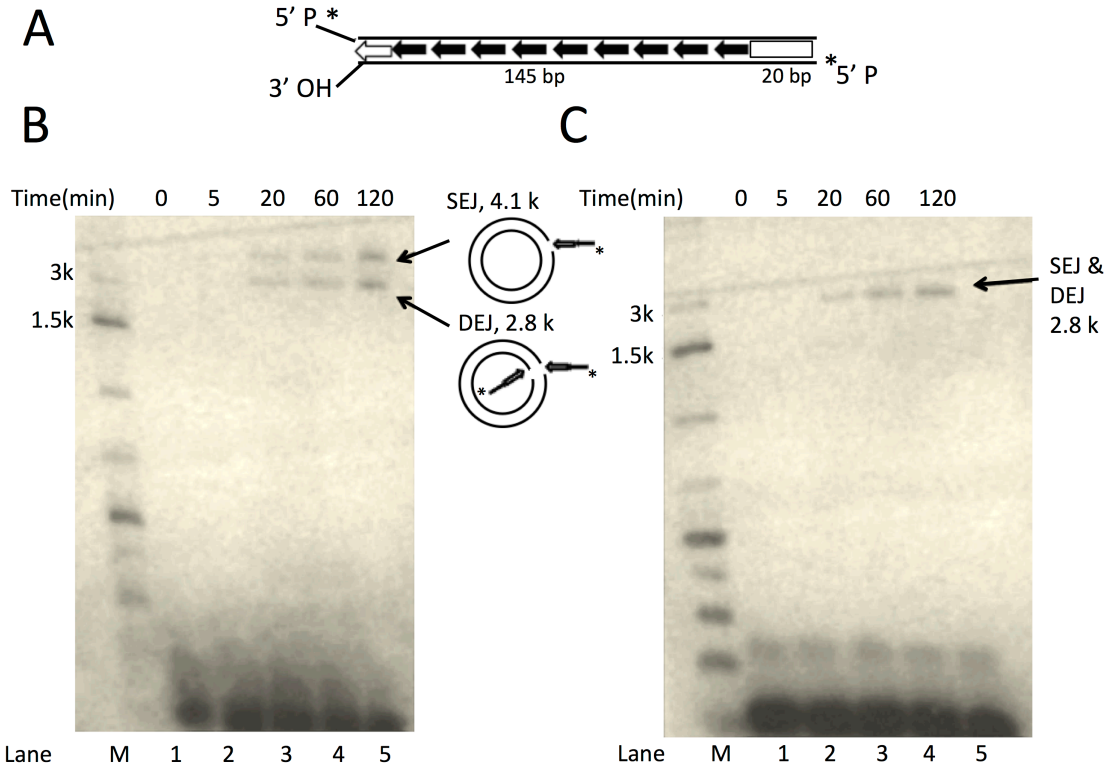

Figure 2-4. *In vitro* analysis of *Muta1* end joining with target DNA.

(A) The 165 bp DNA substrate contains full length *Muta1* L-TIR with 20 bp of *Muta1* internal sequence. Two lines indicate the DNA double strands. On the bottom strand, the 5' P is radiolabeled and the 3' OH is exposed.

(B) Reaction products on a native agarose gel reveal two bands at 4.1 kb and 2.8 kb, which reflect joining product SEJ (nicked circular plasmid formed by joining of one transposon end to one plasmid strand) and DEJ (linearized plasmid formed by concerted joining of two transposon ends to two plasmid strands), as indicated in the diagram on right.

(C) Reaction products on a denaturing agarose gel, only the 2.8 kb is seen, which reflects both DEJ and SEJ produts.

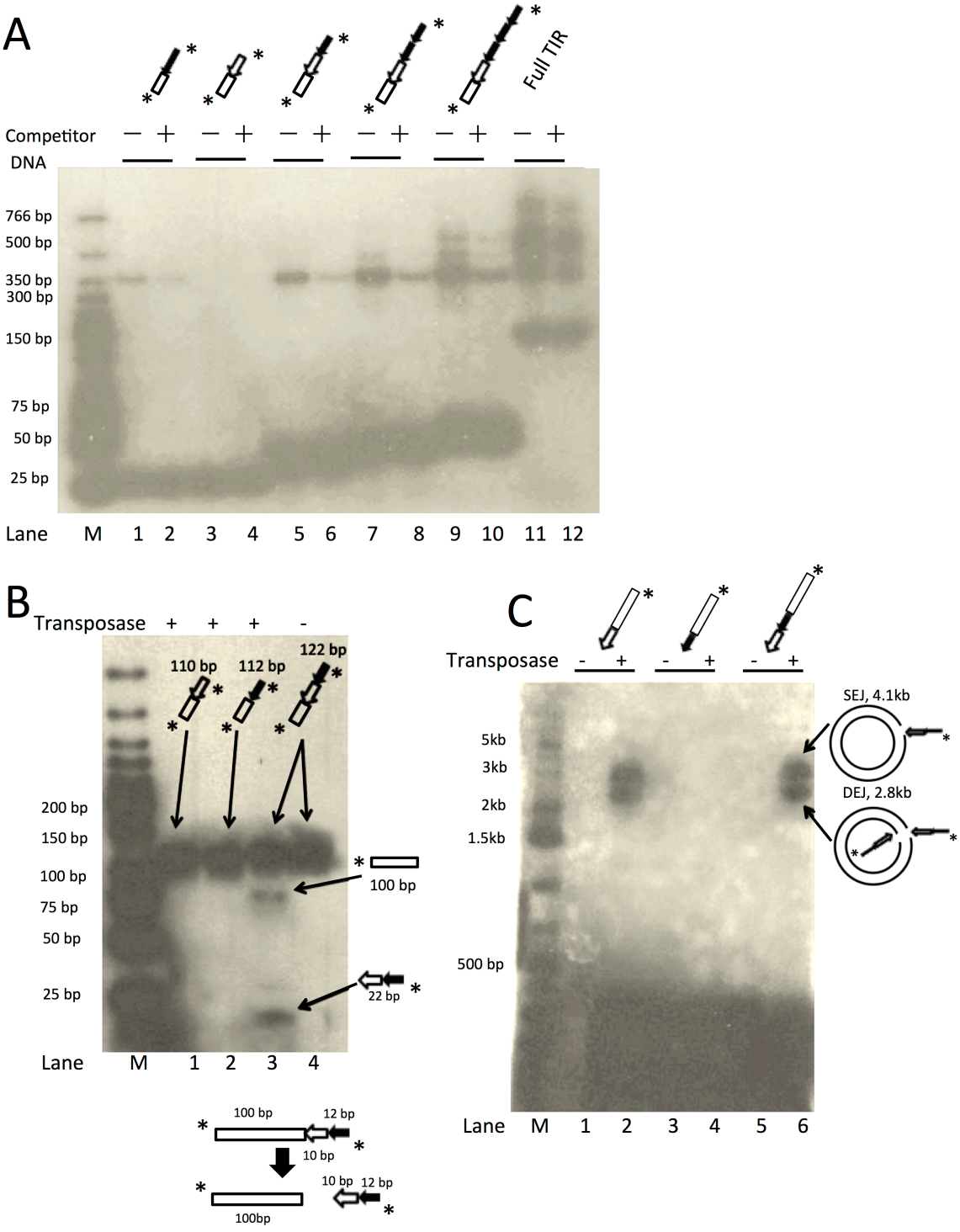
A
Full TIR
Competitor
DNA
766 bp
500 bp
350 bp
300 bp
150 bp
75 bp
50 bp
25 bp
Lane M 1 2 3 4 5 6 7 8 9 10 11 12
B
Transposase + + + -
110 bp
112 bp
122 bp
200 bp
150 bp
100 bp
75 bp
50 bp
25 bp
100 bp
22 bp
Lane M 1 2 3 4
100 bp
12 bp
10 bp
10 bp
12 bp
100bp
C
Transposase - + - + - +
SEJ, 4.1kb
DEJ, 2.8kb
5kb
3kb
2kb
1.5kb
500 bp
Lane M 1 2 3 4 5 6

Figure 2-5. *In vitro* analysis of the impact of the *Muta1* subterminal repeats on tpase binding, double strand break and end joining reactions.

(A) DNA binding assay. Diagrams at top of gel show structure of DNA substrates, all contain 20 bp flanking DNA segment (open box) and subterminal repeats and terminal motif. Unlabeled substrate shown in Figure 2B is added in 100 (+)molar excess as competitor DNA; and 100 molar excess of nonspecific DNA is added in all binding reactions. Reaction products are displayed on native polyacrylamide gel. In each lane, the bottom band indicates the DNA substrate (size from 30 bp to 165 bp), and the top bands (> 300 bp) indicate nucleoprotein complexes.

(B) Donor cleavage assay. Diagram on top shows the structure of DNA substrates, and the bottom diagram shows the predicted structures of cleavage products. Reactions products are displayed on native polyacrylamide gel, the 100 bp band indicates the cleavage product.

(C) End joining assay. Diagram on top shows the structure of DNA substrates, intact pUC19 plasmid is used as target DNA, reaction products are displayed on native agarose gel. The 4.1 kb and 2.8 kb band indicates product of SEJ and DEJ, respectively.

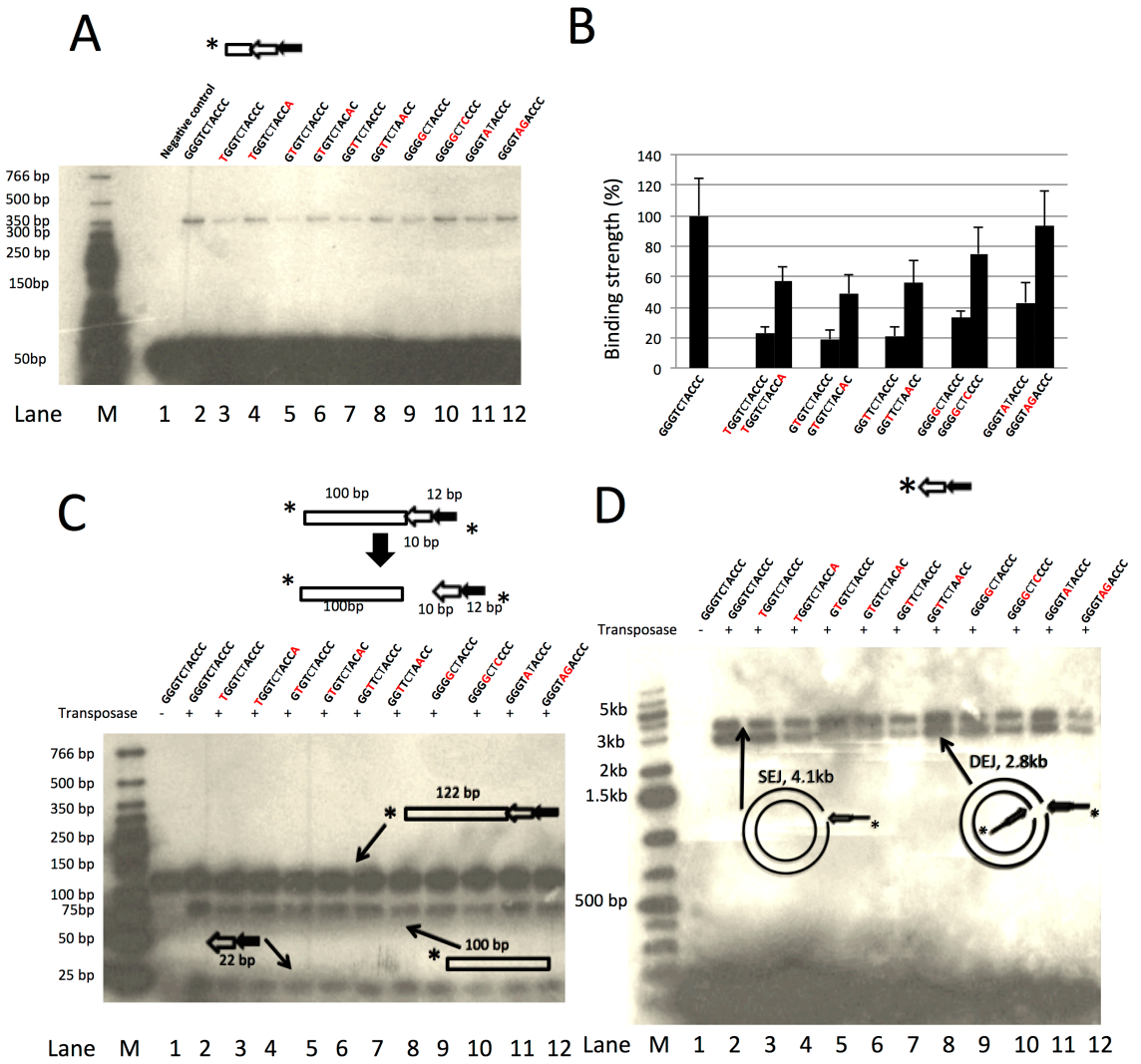
A
Negative control
GGGTCTACCC
TGGTCTACCC
TGGTCTACCA
GTGTCTACCC
GTGTCTACAC
GGTTCTACCC
GGTTCTAACC
GGGGCTACCC
GGGGCTCCCC
GGGTATACCC
GGGTAGACCC
766 bp
500 bp
350 bp
300 bp
250 bp
150bp
50bp
Lane M 1 2 3 4 5 6 7 8 9 10 11 12
B
Binding strength (%)
140
120
100
80
60
40
20
0
C
100 bp
12 bp
10 bp
100bp
10 bp
12 bp
Transposase
766 bp
500 bp
350 bp
250 bp
150 bp
100 bp
75bp
50 bp
25 bp
122 bp
22 bp
100 bp
Lane M 1 2 3 4 5 6 7 8 9 10 11 12
D
Transposase
5kb
3kb
2kb
1.5kb
500 bp
SEJ, 4.1kb
DEJ, 2.8kb
Lane M 1 2 3 4 5 6 7 8 9 10 11 12

Figure 2-6*. In vitro* analysis of the impact of the *Muta1* terminal palindromic motif on DNA binding, double strand break and end joining.

(A) DNA binding assay. Diagram shows the structure of DNA substrates. Red letters indicate mutations in the terminal motif. Reactions products are displayed on native polyacrylamide gel. Unlabeled substrate shown in Figure 2B is added in100 (+) molar excess as competitor DNA; and 100 molar excess of nonspecific DNA is added in all binding reactions. In each lane, the bottom band indicates the DNA substrate (~ 50 bp), and the top band (~ 350 bp) indicates nucleoprotein complexes.

(B) Quantification of the DNA binding of substrates used in (A). The strength of binding was quantified with the ImageJ software and repeated 5 times to generate the standard deviation.

(C) Cleavage assay. Diagram shows the structure of DNA substrates and cleavage products. Sequences of substrates are shown, red letters indicate mutations in terminal motif. Reactions products are displayed on native polyacrylamide gel. In lanes 2-12, the 122 bp and 100 bp bands indicate the DNA substrate and cleavage product, respectively.

(D) End joining assay. Diagram shows the structure of DNA substrates. Red letters indicate mutations in terminal motif. Intact pUC19 plasmid is used as target DNA, and reaction products are displayed on native agarose gel. In lanes 2-12, the 4.1 kb and 2.8 kb bands indicate SEJ and DEJ product, respectively.

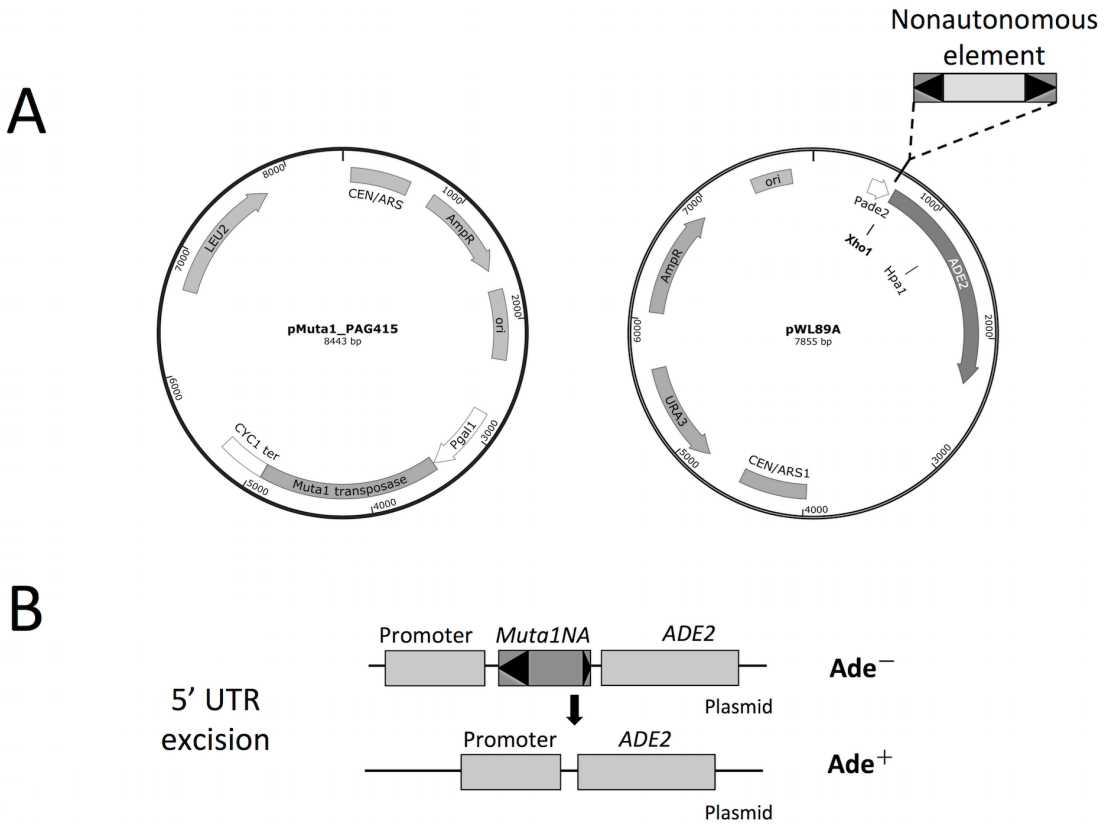

Figure 2-7. Yeast transposition assay constructs.

(A) The structure of transposase expression vector pMuta1_PAG415 and reporter vector pWL89A (see methods). AmpR, ampicillin resistance gene; ori, *E. coli* replication origin; Pgal1, GAL1 promoter; CYC1 ter, terminator; CEN, centromere sequences of yeast chromosomes; ARS, autonomous replication site. Dashed lines indicate the position of nonautonomous elements insertion site in the 5' UTR of *ADE2*.

(B) *Muta1NA1* excision from 5' UTR of *ADE2*.

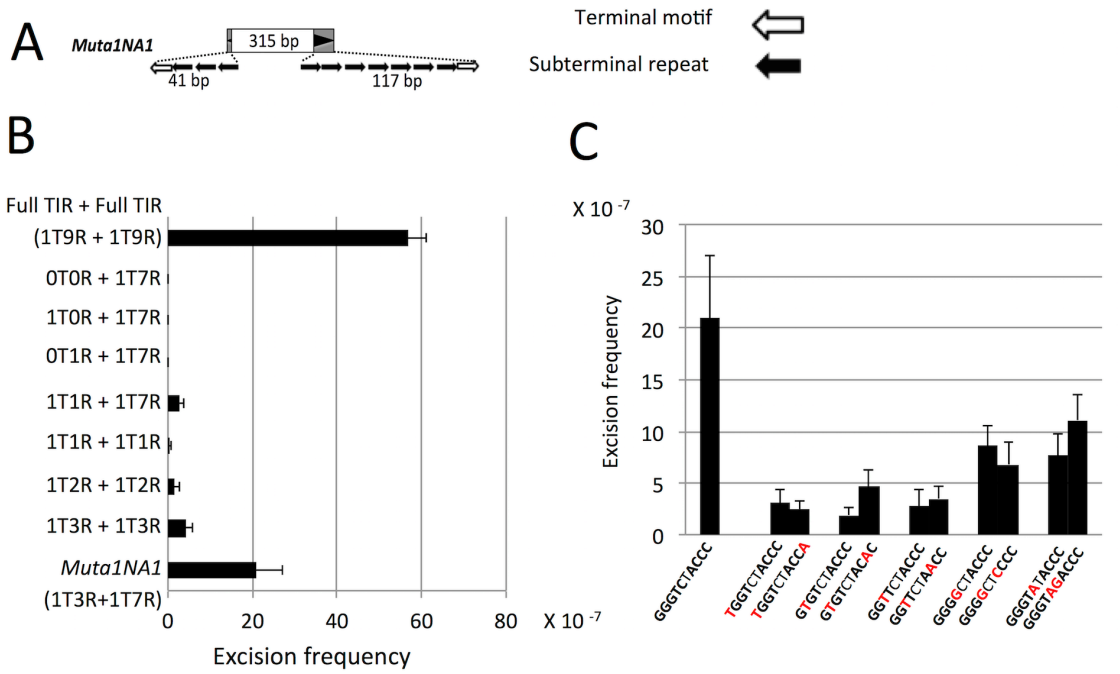

Figure 2-8. Analysis of the impact of *Muta1* subterminal repeats and terminal motif on transposition in yeast.

(A) *Muta1NA1* contains 10bp terminal motif on both ends; 3 copies and 7 copies of the subterminal repeats on left and right end respectively; and a 315bp of internal DNA segment.

(B) Excision frequencies of *Muta1NA1* derivative elements generated by altering the copy number of the two motifs at both ends of *Muta1NA1*. TIR composition of each element is shown at the left (see text for details). For each element, 6 replicates were used.

(C) Excision frequency of mutant versions of *Muta1NA1.* Mutations are introduced at both ends of the terminal motif of *Muta1NA1,* as indicated by red letters. For each element, 6 replicates were used.

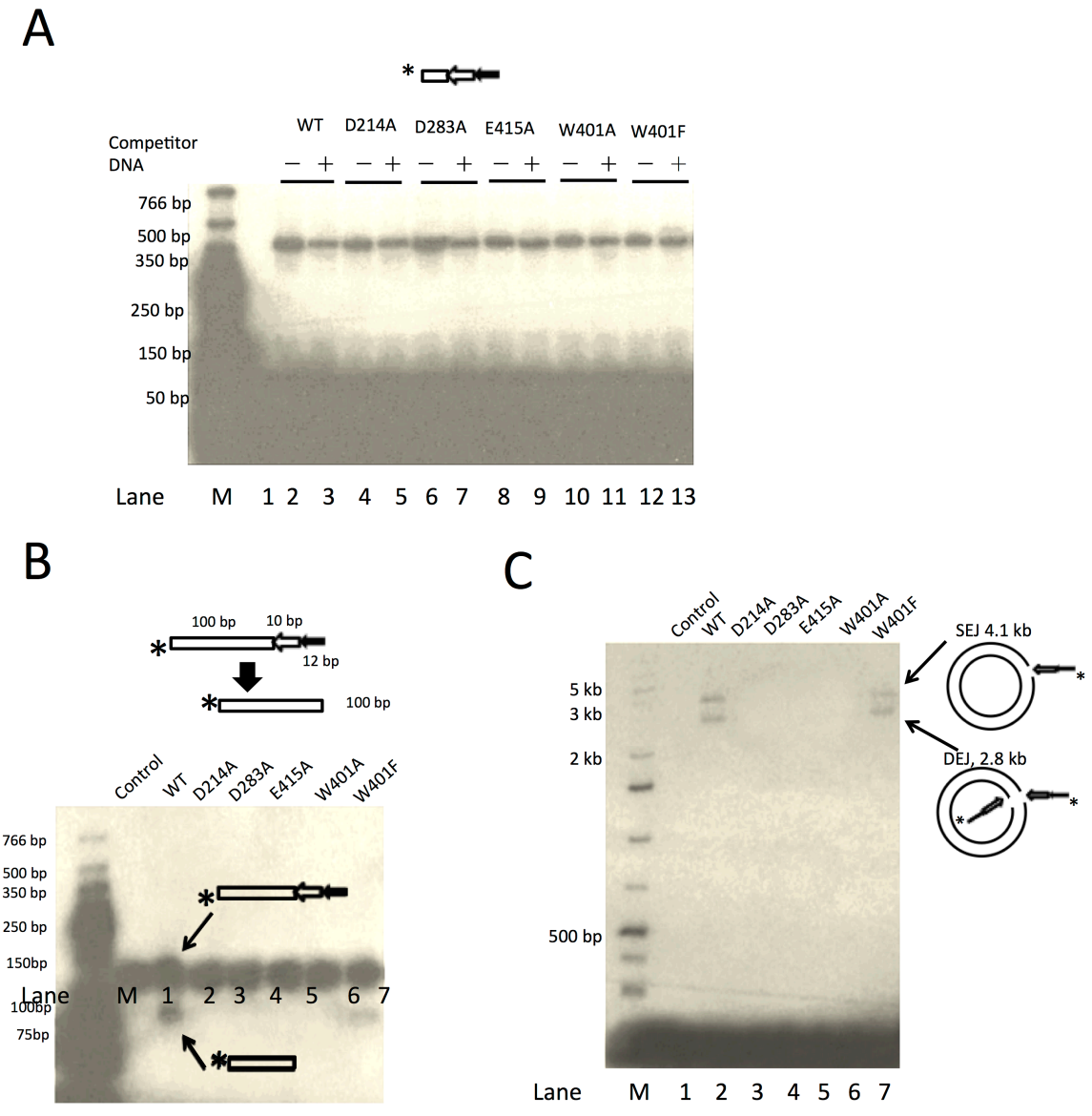
A
WT
D214A
D283A
E415A
W401A
W401F
Competitor DNA
766 bp
500 bp
350 bp
250 bp
150 bp
50 bp
Lane M 1 2 3 4 5 6 7 8 9 10 11 12 13
B
100 bp
10 bp
12 bp
100 bp
Control
WT
D214A
D283A
E415A
W401A
W401F
766 bp
500 bp
350 bp
250 bp
150bp
100bp
75bp
Lane M 1 2 3 4 5 6 7
C
Control
WT
D214A
D283A
E415A
W401A
W401F
SEJ 4.1 kb
DEJ, 2.8 kb
5 kb
3 kb
2 kb
500 bp
Lane M 1 2 3 4 5 6 7

Figure 2-9. *In vitro* mutagenesis analysis of the role of conserved residues of *Muta1* transposase in transposition.

(A) DNA binding assay using wild type and mutant transposases. Diagram shows the structure of DNA substrate. Unlabeled substrate shown in Figure 2B is added in 100 (+) molar excess as competitor DNA; and 100 molar excess of nonspecific DNA is added in all binding reactions. Binding products are displayed on native polyacrylamide gel. In each lane, the bottom band indicates the DNA substrate (~ 50 bp), and the top band (~ 350 bp) indicates nucleoprotein complexes.

(B) Cleavage assay using wild type and mutant transposases, diagram shows the structure of DNA substrates and predicted cleavage products. Reactions products are displayed on native polyacrylamide gel, the 120bp and 100 bp bands indicate DNA substrates and cleavage product, respectively.

(C) End joining assay using wild type and mutant transposases, diagram shows the structure of DNA substrates. Intact pUC19 plasmid is used as target DNA. Reaction products are displayed on a native agarose gel, the 4.1 kb and 2.8 kb bands indicate products of SEJ and DEJ, respectively.

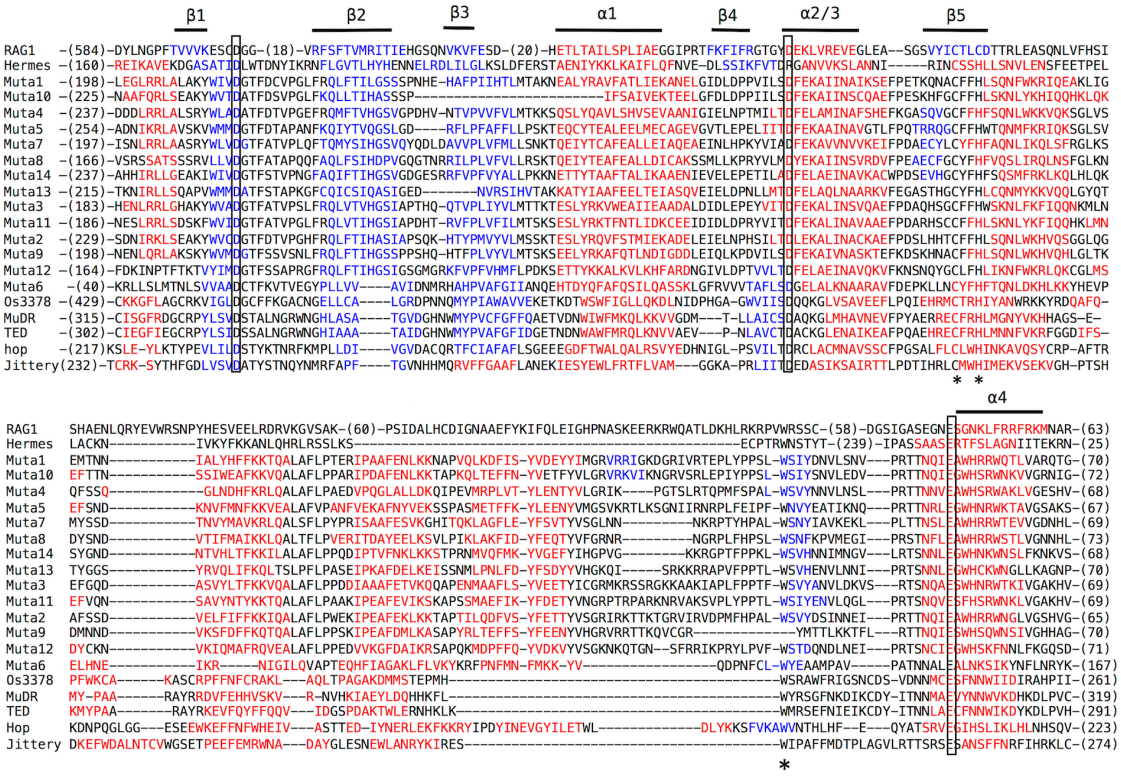

Figure 2-10. Comparison of sequences and predicted secondary structures of transposases and RAG1 recombinase.

DDE domain sequences of *Muta1-14, MuDR, TED, Os3378, Jittery, Hop, Hermes* and RAG1 were aligned with ClustalW (all MULE sequences) and by eye (*Hermes* and RAG1). The secondary structure features are numbered as for RAG1. The DDE triad is boxed; the conserved CxxH motif and W residue are labeled by asterisk. Red letters represent predicted α-helix and blue letters indicate predicted β-strand.

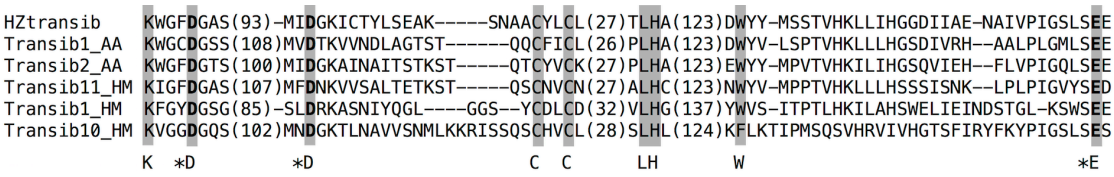

Figure 2-11. Multiple alignment of the DDE domain in *Transib* transposases.

Conserved residues are highlighted, asterisks indicate the DDE triad, the W residue is further discussed in the text. Distances between the conserved blocks are indicated in the number of amino acid residues.

Chapter 3 Characterization of the Pack-*Mutator*–like transposable elements (Pack-MULEs) in the mosquito *Aedes aegypti*

Abstract

Gene duplication followed by sequence and functional divergence is an important mechanism for evolution of new genes. Pack-MULEs, nonautonomous *Mutator*-like transposable elements (MULEs) that carry gene or gene fragments, are able to duplicate and amplify gene fragments on a large scale in plants, including maize, rice and *Arabidopsis*. Prior studies have shown that Pack-MULEs are potentially involved in generating new open reading frames and regulating host gene expression. However, despite their abundance and importance, Pack-MULEs have not been reported outside the plant kingdom.

Here, we report the identification of 1,378 Pack-MULEs in the genome of the mosquito, *Aedes aegypti,* including 663 chimeric elements, which acquired sequences from two or more genomic locations. *A. aegypti* MULEs preferentially acquire fragments from genic regions, with 423 host genes involved. Among the 2,249 gene fragments identified in Pack-MULEs, only 83 (2.8%) are acquired from coding sequences, while the rest are from introns or UTRs. Of the *A. aegypti* Pack-MULES, 9.7% are expressed, 11.8% directly generate sRNAs, and no evidence of the translation of Pack-MULE sequences could be detected. Comparison of the

sequences acquired by the host genes indicates that fragments of genomic DNA have been captured, amplified and rearranged on a large scale and for a long period of time in the *A. aegypti* genome. Gene acquisition activity of Pack-MULEs might provide an important mechanism for the evolution of genes in *A. aegypti*.

Introduction

Transposable elments (TEs) are DNA sequences that can move from one genomic location to another. In the process, they can generate genomic alterations including insertions, excisions, duplications and translocations (Finnegan, 1989; Wicker et al., 2007). With this ability and their ubiquitous presence in eukaryotic genomes, TEs are considered to play important roles in genome evolution (Yang et al., 2006). TEs are classified as class 1, which utilize an RNA intermediate in their movement and class 2, which use a "cut and paste" mechanism where the transposition intermediate is the element itself. An exception for class 2 TEs is *Helitron,* which transposes via a rolling-circle mechanism (Kapitonov et al., 2001).

In addition to occasionally duplicating during transposition, several types of TEs are found to capture and duplicate host genes or gene fragments, in which the sequence and functional divergence provides means for the generation of new genes (Bureau et al., 1994; Flagel et al., 2009). In a process called transduction, class 1 TEs can acquire flanking host genomic sequences via transcriptional readthrough (Moran et al., 1999). For example, the human *L1* element has been shown to transduce adjacent genes and other genomic sequences on a large scale (10% of *L1* elements are involved in transduction, Moran et al., 1999). Several class 2 TE superfamilies have also been reported to acquire fragments of host genes, including MULE*, CACTA* and *Helitron* (Pickeral et al., 2000; Jiang et al., 2004; Morgante et al., 2005). The mechanism of gene fragment acquisition by class 2 TEs is unclear, but the capture of

introns indicates that the process occurs at the DNA level, which differs from transduction (Jiang et al., 2004). Duplicated gene fragments may have functional impacts on the host depending on whether they are transcribed, translated or associated with small RNAs (sRNAs).

Mutator–like transposable elements (MULEs), a class 2 TE superfamily that was first identified in maize (Robertson 1978; Bennetzen et al., 1984). Subsequent research indicated that this superfamily is widespread in plants, animals, fungi, protozoans, and are closely related to the prokaryotic IS*256* family (Eisen et al., 1994; Chalvet et al., 2004; Neuveglise et al., 2005; Pritham et at., 2005). They are characterized by long terminal inverted repeats (TIRs, >100 bp) and 9–11 bp target site duplication (TSD). MULE families typically consist of a few autonomous elements containing the transposase gene, and up to thousands of nonautonomous elements that can be mobilized in trans by the transposase produced by the autonomous copies (Lisch 2002). In a MULE family, all members share the TIR sequence, but many nonautonomous MULEs are not simple deletion derivatives of their coreresponding autonomous copies. Instead, they carry diverse internal sequences between the TIRs including host gene fragments (Bennetzen et al., 1994).

MULEs carrying host gene or gene fragments are called Pack-MULEs (Jiang et al., 2004). They have been identified in other plant species, including maize and *Arabidopsis* (Carels et al., 2000; Wong et at., 2001). Pack-MULEs have been reported

to duplicate host gene fragments on a large scale, especially in rice, where more than 2,900 Pack-MULEs were found to carry fragments acquired from over 1,500 genes (Jiang et al., 2011; Ferguson et al., 2013). Rice Pack-MULEs contribute to the genome architecture and evolution through modifying the GC content and GC gradient in the genome (Ferguson et al., 2013; Juretic et al., 2011). In addition, of the rice Pack-MULEs, over 20% are transcribed; more than half are associated with sRNAs, and a few are translated (Juretic et al., 2011; Hanada et al., 2009).

A prior study (see chapter1) identified 14 MULE families in the genome of the mosquito, *Aedes aegypti,* which can spread dengue fever, yellow fever, chikungunya, zika, and many other diseases (Womack M 1993; Marchette et al., 1969; Nene et al., 2007). The 14 MULE familie all have 8-9 bp TSD, 12 of 14 have TIRs >100 bp and two have TIRs <50 bp. Ten of the 12 long TIR families contain short subterminal tandem repeats (9-15 bp). Most families also contain derivative nonautonomous elements, which share high sequence similarity in their TIR and subterminal regions, but carry heterogeneous internal sequences including fragments of host genes (see chapter 1). Here we report the identification and characterization of Pack-in the *A. aegypti* genome.

Results

The Inventory of MULEs in the *A. aegypti* Genome

A prior study identified Pack-MULEs in the *A. aegypti* genome (see chapter 1), to identify and characterize Pack-MULEs, we first assessed the abundance of MULEs in the *A. aegypti* genome. Repeat families with TIR and 8-9 TSD (which is a feature of MULEs) in *A. aegypti* were identified with RECON as previously described (Jiang et al., 2004; Hanada et al., 2009; Ferguson et al., 2013), and the resulting families were examined for and 8-10 bp TSD, which is the feature of MULEs. Because some of the MULE families identified previously have unusually short TIRs (as short as 16 bp, see chapter 1), we considered all TE families with 8-9 bp TSD as potential MULEs, regardless of the TIR length. To achieve accurate classification, the repeat families were compared to the 14 *A. aegypti* MULE families (see chapter 1) and all MULE families in Repbase libraries. A TE family with 8-10 bp TSD was defined as a MULE only if (1) it shares sequence homology to MULE TIRs (e- value $<10^{-10}$); or (2) the internal sequences of family members share significant similarity (e- value $<10^{-10}$) to the MULE transposases. Our search resulted in 24 MULE families (with 2,235 elements) and 53 families (with 6,394 elements) that cannot be classified because the lack of sequence similarity to known MULEs and thus excluded in the following analysis. Prior study in rice indicated that MULEs with long TIR were more often involved in gene acquisition events (Ferguson et al., 2013), to test if this finding is also true for *A. aegypti* MULEs, the 2,235 MULEs with TSDs were sorted into 19 long-TIR MULE families (TIR > 50 bp) and 5 short-TIR MULE families (TIR < 50 bp)

based on the similarity and length of TIRs. Among the MULEs, 1,631 (72.9%) were long-TIR MULEs, indicating this type is more predominant than the short-TIR MULEs (Figure 3-1).

Regarding their insertion sites in the genome, 1,005 (40.7%) MULEs are located in gene bodies or within 5kb upstream or downstream of genes; in a control dataset of equal number of randomly selected genomic sites (Figure 3-2 A&B, see methods), indicating *A. aegypti* MULEs may preferentially target genic regions for insertion, as previously reported for plant MULEs and many other TE superfamilies (Lisch 2002).

Identification of Pack-MULEs in the *A. aegypti* Genome

Based on the Pack-MULE identification protocol described previously (Jiang et al., 2004; Hanada et al., 2009; Ferguson et al., 2013), the 2,235 *A. aegypti* MULEs were used in BLAST searches to determine if they were Pack-MULEs. If the internal sequence of a MULE has a significant match (e- value $<10^{-10}$) to a non-TE genomic homolog, the genomic homolog is called the host fragment or host gene (if it is in a genic region). As a result, 1,378 MULEs were found to contain host genic fragments, and 304 MULEs contain intergenic fragments. In addition, 62.6% of the long-TIR MULEs and 58.9% of the short-TIR MULEs contain host fragments (Figure 3-1). According to the internal sequence contained within the TIR, MULEs were further sorted into three groups (Figure 3-2 C&D): (1) Pack-MULEs, as defined previously, refer to elements containing sequences matching functionally annotated genes or

hypothetical genes; (2) MULE-Tpase, refers to elements containing sequences with significant homology (e- value $<10^{-10}$) to known MULE transposases; and (3) MULE, refers to elements without any identifiable host sequences. To test if these *A. aegypti* Pack-MULEs are an artifact of sequence assembly, 30 were randomly selected and sequenced after PCR amplification from isolated genomic DNA. The results confirmed that all virtual elements exist in the genome and that they carry gene fragments and are flanked by 8-9 bp TSDs.

Among the Pack-MULEs, 672 carry a single gene fragment, 411 carry two gene fragments and 295 carry three or more fragments (ranging from 3 to 8). A total number of 2,641 gene fragments were identified; their average length is 178 bp (ranging from 39 bp to 1,724 bp) and the average sequence identity to host genes is 81.9% (ranging from 63.7% to 100%, Figure 3-3). Thirty-eight acquired fragments share high sequence identity with host genes (> 98%), indicating possible ongoing gene acquisition activity in the *A. aegypti* genome.

Pack-MULEs containing the same acquired fragments were clustered as subfamilies, the results showed that 1,211 out of 1,378 Pack-MULEs are single-copy elements, and the remainders are multi-copy. The multi-copy subfamilies presumably arose by replicative transposition of an ancestral Pack-MULE that had previously acquired gene fragments. 51 subfamilies have two to four copies; and 6 subfamilies have five or more copies (Table 3-1). These multi-copy subfamilies were generated by

transposition instead of other duplication events because all subfamily members have distinct TSDs. The element sequence identity within each multi-copy subfamily ranges from 91% to 95%, suggesting they may not be involved in recent transposition events. The average sequence identity between the acquired fragments in these multi-copy Pack-MULEs and the host copies is 89.4% (ranging from 79.3% to 91.9%, Table 3-1), suggesting that the multi-copy Pack-MULEs may not be involved in recent acquisition events. Notably the Pack-MULEs of higher copy subfamilies (copy number >4) are shorter than the single copy elements (Table 3-1).

To study the timing of acquisition events, all acquired gene fragments in each of the Pack-MULE families were classified based on their sequence identity to the parental sequences. For the four families containing more than 100 Pack-MULEs, analysis of the divergence between acquired fragments and their host copies produce a temporal pattern indicative of successive gene acquisition activity (Figure 3-4).

Structures of representative Pack-MULEs are shown (Figure 3-5), which all carry different internal sequences. The 423 host genes identified from the 1,378 Pack-MULEs involve diverse cellular processes including transcription, signal transduction and cell metabolism, indicating that a wide variety of gene fragments have been captured by Pack-MULEs. From the evolutionary standpoint, of great interest is the discovery of Pack-MULEs that are involved in sequential acquisition events, which not only have potential to create novel genes through rearrangement

and fusion of diverse genomic sequences, but also provide hints for the study of the mechanism of gene acquisition. A total number of four step-wise acquisition events were identified, in which a fragment (fragment A) is first captured by Pack-MULE, and next a second fragment (fragment B) is inserted into fragment A (Figure 3-5C). All of the four Pack-MULEs involved in sequential acquisition are single-copy element, their length varies from 3.2 kb to 5.6kb (mean is 4.4 kb). For the four Pack-MULEs, the average length and sequence identity of fragment A and fragment B is 508 bp (ranging from 106 bp to 1,175 bp) and 85.9% (ranging from 78.3% to 92.5%), and 253 bp (ranging from 52 bp to 487 bp) and 93.4% (ranging from 85.4% to 98.2%), respectively. However, in all four cases, the intermediate element of the sequential events (a Pack-MULE containing fragment A but not fragment B) is not identified.

A. aegypti Pack-MULEs preferentially acquire genic sequences

Because plant Pack-MULEs preferentially acquire DNA fragments from genic regions, we asked if *A. aegypti* Pack-MULEs have the same preference preference (Ferguson et al., 2013). To address this question, the host copies of acquired fragments in Pack-MULEs were examined. Among the 3,288 host fragments, 2,249 are captured from genic regions and 1,039 from intergenic regions. Considering genic sequences (241.55 Mb) account for roughly 17.5% of the *A. aegypti* genome, and 68.3% of the acquired sequences, this result indicates that genic sequences are overrepresented as the source of the acquired fragments (Figure 3-6A, $P < 0.05$, χ^2

test). For the fragments captured from genes, we further analyzed their sequence oringin and identified 83 fragments from coding sequences, 1,841 introns, 134 5' UTRs and 191 3' UTRs, this could be explained by the long introns of *A. aegypti*, which account for about 84.4% of *A. aegypti* gene space preference (Ferguson et al., 2013). In addition, among different age groups (defined by the sequence identity to the host copy), the distribution of fragments acquired from each genetic region in not significantly different (Figure 3-6B, P >0.5, χ^2 test).

The junction of acquisition events in Pack-MULEs

Relatively new acquisition events (>95% sequence identity between acquired fragments and host sequences) permitted the identification of acquisition breakpoints and boundary of acquired fragments in Pack-MULEs. These junction sequences allowed us to study the features of gene acquisition events, for example, if there is sequence homology between the MULE internal sequence flanking the acquisition junction and the ends of acquired sequence. To this end, the boundary of 145 acquired fragments were identified and analyzed.

After obtaining the junctions, for each acquired fragment, three sets of sequences were compared: (1) 30 bp border of the fragment; (2) 30 bp Pack-MULEs sequences flanking the fragment insertion site; and (3) 30 bp host gene sequences flanking the host copy of the fragment. However, in most cases, no apparent sequence homology between the boundary of acquired fragments and their surrounding sequences were

identified. To evaluate GC content plays a role in gene acquisition process, 100 bp of Pack-MULE and host gene sequences flanking the acquired fragment or its host copy were divided into ten 10 bp bins (Figure 3-7A). The result revealed that the acquired fragments were often located in Pack-MULE internal regions with lower GC content (< 30%, $P < 0.1$, χ^2 test), whereas the GC content of the sequences surrounding the host fragments was not significantly different from other regions in the host gene (Figure 3-7A, $P > 0.5$, χ^2 test). In addition, the GC content of the entire Pack-MULE internal region, the entire host genes, acquired fragments and host fragments were not significantly different and no apparent GC gradient was observed (Figure 3-7 B&C, $P > 0.5$, χ^2 test), *A. aegypti* Pack-MULEs do not specifically acquire sequences with higher GC content, different from rice Pack-MULEs (Hanada et al., 2009; Ferguson et al., 2013).

Expression of *A. aegypti* Pack-MULEs

To explore their possible functionality, we examined the transcription profile of the Pack-MULEs (the entire elements) was examined using RNA-Seq data from 42 *A. aegypti* expression libraries representing different mosquito developmental stages and diverse tissues (Akbari et al., 2013). A Pack-MULE was considered transcribed only if its FPKM (fragments per kilobase of exon per one million fragments mapped) value of unique mapped reads was at least 1.0 in at least one RNA-Seq library. Only 119 (8.6%) Pack-MULEs are transcribed. To test if the small portion of transcribed Pack-MULEs is because of low coverage of the RNA-Seq libraries, transcription of

host genes was examined using the same method. Of the 423 host genes, 75.2% (318) were transcribed (FPKM >1 in at least one of the 42 libraries).

As another measure of transcription, all Pack-MULEs were used to query a database of 307,080 unique *A. aegypti* cDNAs and ESTs. Only 99 (7.2%) Pack-MULEs had cDNA/EST matches (sequence identity > 99.5%), whereas 278 (65.7%) host genes had matches. This indicates that most Pack-MULEs are not transcribed, consistent with the RNA-seq data. After combining the results of two approaches, 134 transcribed Pack-MULEs were identified, representing 9.7% of total Pack-MULEs.

Among the 134 transcripts produced from Pack-MULEs, two contain fragments captured from coding sequences, 19 contain UTR fragments while the rest contain either intron or intergenic fragments. Two of the transcribed Pack-MULEs are multi-copy elements and 68 out of 134 (50.7%) transcribed Pack-MULEs are chimeric elements, similar to the proportion of chimeric elements in total Pack-MULEs (48.1%). Examples of transcribed Pack-MULEs are shown (Figure 3-8).

To determine if any Pack-MULEs were translated, peptide sequences identified from proteomic analyses in *A. aegypti* were collected (Hugo et al., 2013; Marinotti et al., 2014; Nunes et al., 2016; Oktarianti et al., 2015; Sirot et al., 2003; Popova-Butler et al., 2009; Almeras et al., 2010; Bayyareddy et al., 2012). The 2,731 peptides represent 16.4% of the annotated genes based on the AaegL3.3 annotation

(VectorBase). However, none of the Pack-MULEs had exact match with the peptide sequences, whereas, 62 out of 423 (14.6%) host genes had unique and perfect matches.

Small RNA production of *A. aegypti* Pack-MULEs

Plant MULEs were shown to be important sources of plant sRNAs production (Hanada et al., 2009; Ferguson et al., 2013). To test if *A. aegypti* Pack-MULEs also generate sRNAs, 4 *A. aegypti* sRNA libraries representing different developmental stages were searched (Aguiar et al., 2015; Arensburger et al., 2011). We first tried to identify sRNAs associated with the TIR of MULEs because the inverted repeats, if transcribed, could trigger sRNA biogenesis. Perfectly matched sRNAs were identified in all 19 long-TIR MULE families and 2 of the 5 short-TIR families in at least one library. Next, we looked for sRNAs derived from the entire internal region of Pack-MULEs. For easier comparison, the sRNAs reads mapped to Pack-MULEs were classified into two categories: (1) self-sRNAs are those with unique exact matches with Pack-MULEs; and (2) shared-sRNAs are those perfectly mapped to both Pack-MULEs and their host genes, meaning they are generated from either Pack-MULEs or the host genes. Only 11.8% (163) of the Pack-MULEs have self-sRNAs; 6.1% (84) share sRNAs with host genes; whereas 65.9% (908) are not associated with sRNAs. In contrast, 32.4% of host genes have unique mapped sRNAs, suggesting the low level of sRNA mapped to Pack-MULEs is not caused by the quality or low coverage of sRNA libraries.

Discussion

Pack-MULEs have been shown to duplicate and rearrange host gene fragments on a large scale in plant genomes, thereby might provide an important mechanism for genome evolution. In this study, for the first time, we report the identification and characterization of Pack-MULEs in a non-plant species, the mosquito, *A. aegypti*. In the *A. aegypti* genome53 families of over 6,000 individual elements were excluded in our study because of the ambiguity in classification, although some of them may be MULEs. Therefore, our identification of 2,235 MULEs and 1,378 Pack-MULEs is likely an underestimation. Both the *A. aegypti* and rice genomes contain large numbers of Pack-MULEs (Table 3-2), which have similar structural features; thus comparison of Pack-MULEs in the two species may be informative. The shared features between *A. aegypti* and rice Pack-MULEs include the predominance of long-TIR elements (Figure 1), preference of insertion into gene rich regions (Figure S1), and preference of acquiring genic sequences (Figure S3A). This preference is not the result of shorter retention time of intergenic fragments, as the number of relatively recent intergenic acquisition events is not significantly higher than old acquisitions events (Figure S3B, 28). In both genomes, the acquired fragments in some Pack-MULEs share high sequence identity with host genes (up to 100%), indicating possible ongoing gene acquisition activity in both genomes. In addition, most Pack-MULEs in *A. aegypti* and rice are single-copy elements (Table 3-1, Ferguson et al., 2013).

Pack-MULEs represent 61.6% and 21.2% of the MULEs in *A. aegypti* and rice, respectively. Forty-eight percent of *A. aegypti* Pack-MULEs contain fragments acquired from two or more genomic locus, whereas 22.4% of rice Pack-MULEs are chimeric elements (Table 3-2), this suggests that the *A. aegypti* Pack-MULEs may have higher gene acquisition activity than rice Pack-MULEs. Among the acquired fragments of *A. aegypti* Pack-MULEs, only 2.8% were captured from protein coding sequences, whereas the proportion is 66.5% for rice Pack-MULEs (Table 3-2) (Ferguson et al., 2013). This could be explained by the long introns of *A. aegypti*, which account for about 84.4% of *A. aegypti* gene space, whereas coding sequences account for 48.6% of the gene space in rice. A smaller proportion of *A. aegypti* Pack-MULEs are transcribed or are the source of sRNAs compare to rice Pack-MULEs (Table 3-2). One of the most notable contribution of rice Pack-MLUEs is that they can reshape the GC content of the genome (Ferguson et al., 2013), however, this phenomenon is not observed for *Ae. aegypti* Pack-MULEs.

In 145 relatively recently acquired fragments (>95% sequence identity to host sequences) by *A. aegypti* Pack-MULEs, little or no sequence homology was detected between and the Pack-MULE and host gene sequences flanking the acquired fragments. Although *A. aegypti* Pack-MULEs have no preference of acquiring GC rich sequences (Figure S3B&C), different from rice Pack-MULEs (Ferguson et al., 2013), the acquisition break points are located in low GC regions within Pack-MULEs (Figure S3A), suggesting GC content is important for the acquisition process, as

hypothesized for rice Pack-MULEs. Moreover, 62.6% of the long-TIR MULEs and 58.9% of the short-TIR MULEs contain acquired fragments (Figure 1), suggesting that TIR length is not likely to be important for gene acquisition.

Conclusions

Our study reveals that *A. aegypti* Pack-MULE mediate the duplication of host gene fragments on a large scale, including 1,378 Pack-MULEs, 2,641 acquired fragments and 423 host genes. *A. aegypti* Pack-MULEs preferentially insert into gene rich regions and prefer to acquire sequences from genic regions. Although only a small proportion of *A. aegypti* Pack-MULEs are expressed and associated with sRNAs, they are involved in the acquisition, amplification and rearrangement of host gene fragments on a large scale, and providing means for the genome evolution.

Methods

Identification of MULEs and Pack-MULEs

First, repeat families (family members sharing >90% TIR sequence identity) with TIRs (>10 bp) were identified from the *A. aegypti* genome (AaegL3 build, www.vectorbase.org/organisms/aedes-aegypti/liverpool/aaegl3) with RECON as described (Bao et al., 2002). The 74 resulting families (< 30 kb) with 8-10 bp TSD were selected for further analysis, and the TSD of each element was manually verified. A family is considered as MULE if its members share significant homology (e- value $<10^{-10}$) with known MULEs (records in RepBase, 36)in the TIR sequence, or the internal sequences of family members share significant homology (e- value $<10^{-10}$) with known MULE transposases.

One hundred bp flanking DNA sequences of each MULE were retrieved and used for BLASTN searches against the *A. aegypti* genome (AaegL3 build) and compared to the genome annotation (release AaegL3.3, www.vectorbase.org/organisms/aedes-aegypti/liverpool/aaegl3) to determine adjacent genes at each insertion site. For the control data set, 1,631 and 604 genome coordinates across the 4,757 scaffolds were randomly generated to represent random insertions of the long-TIR and short-TIR MULEs, respectively. Coordinates were compared to the genome annotation (release AaegL3.3) to determine the surrounding sequences and genes. Random insertion sites generation used 1,000 replicates to estimate the expected number of insertions (and standard deviations) in each category.

Annotation of Pack-MULEs was similar to that described previously with modifications (Jiang et al., 2004; Hanada et al., 2009). The internal sequence of each MULE was used in BLAST searches against the *A. aegypti* genome (AaegL3 build) and the genome annotation (release AaegL3.3) as previously described (Jiang et al., 2004; Hanada et al., 2009), the hit sequence with the highest similarity score that was not associated with MULE TIRs was considered as the host copy of the MULE internal sequence. MULEs were sorted into Pack-MULEs, MULE-Hypo, MULE-Tpase, MULE-intergenic and non-Pack-MULE based on the origin of their internal sequences (see text).

Identification of acquisition junctions of Pack-MULEs

Pack-MULEs containing captured fragments with greater than 97% sequence identity to host sequences or involved in sequential acquisition events were selected to determine the boundary of the alignment of acquired fragments and parental copy. After manual verification of the alignement, the corresponding junction sequences of each acquired fragment were used for further analysis (see text).

GC content analysis

To calculate the GC content of Pack-MULEs and non-Pack-MULEs, first the nested TE insertions were removed, then the TIR sequences (on both ends) and the internal regions were divided into 2 and 10 equal-sized bins, respectively. GC content of each bin was calculated using the biopiece tool set (www.biopieces.org). To determine

the GC content of acquired fragments and their host copies, each fragment was divided into 10 equal-sized bins, and used for GC content calculation by the biopiece tool.

Expression analysis

In the RNA-seq approach, sequencing datasets from relevant reference (Akbari et al., 2013) were downloaded (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP026319), the reads were aligned to the *A. aegypti* genome (AaegL3 build) using Bowtie (Langmead et al., 2012). In the alignment, up to one mismatch and two reportable alignments was allowed. Pack-MULEs or genes were considered as transcribed if the FPKM value of unique mapped reads is greater than 1 in at least one of the 42 expression libraries.

In the cDNA/EST approach, the expression dataset (https://www.vectorbase.org/download/aedes-aegyptiest-clipped2012-12fagz) was downloaded and compared to all *A. aegypti* Pack-MULEs and genes through BLASTN. A Pack-MULE or gene is classified as transcribed if it had a cDNA or EST match with >99.5% sequence identity.

2,731 peptide sequences identified from the relevant references (Hugo et al., 2013; Marinotti et al., 2014; Nunes et al., 2016; Oktarianti et al., 2015; Sirot et al., 2003; Popova-Butler et al., 2009; Almeras et al., 2010; Bayyareddy et al., 2012) were

downloaded and used in TBLATN search against all *A. aegypti* Pack-MULEs and genes. A Pack-MULE or gene was considered as translated on if a perfect peptide hit was detected.

Small RNA analysis

Small RNA sequencing datasets were downloaded (https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP034669, https://www.ncbi.nlm.nih.gov/sra/SRX877435[accn]), and the reads were aligned to the *A. aegypti* genome (AaegL3 build) using Bowtie. In the alignment, no mismatch was allowed. The mapped sRNAs were sorted into self-sRNAs, shared-sRNAs and other sRNAs based on the their origin (see text).

Figures and tables

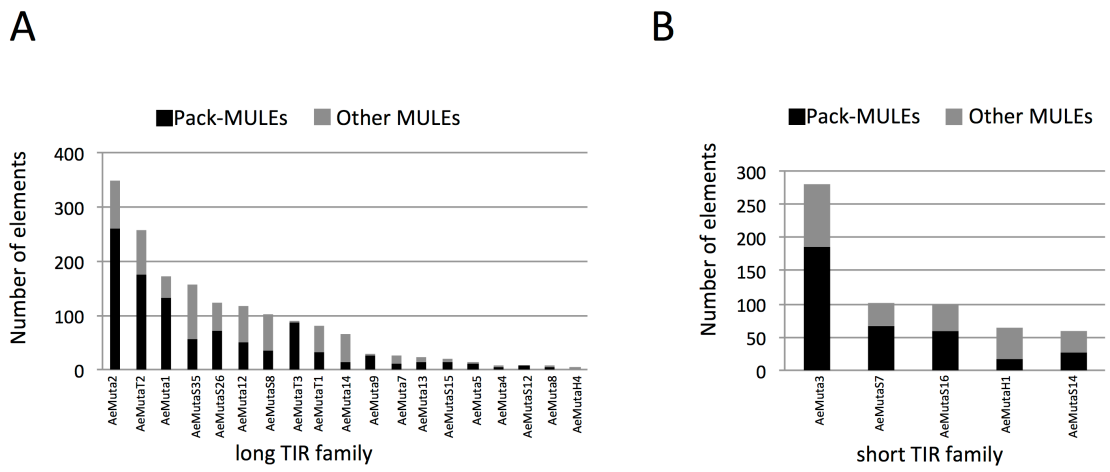

Figure 3-1. Partition of Pack-MULEs among long-TIR and short-TIR MULE families in the *A. aegypti* genome.

(A) Copy number of Pack-MULEs in long-TIR MULE families.

(B) Copy number of Pack-MULE in short-TIR MULE families.

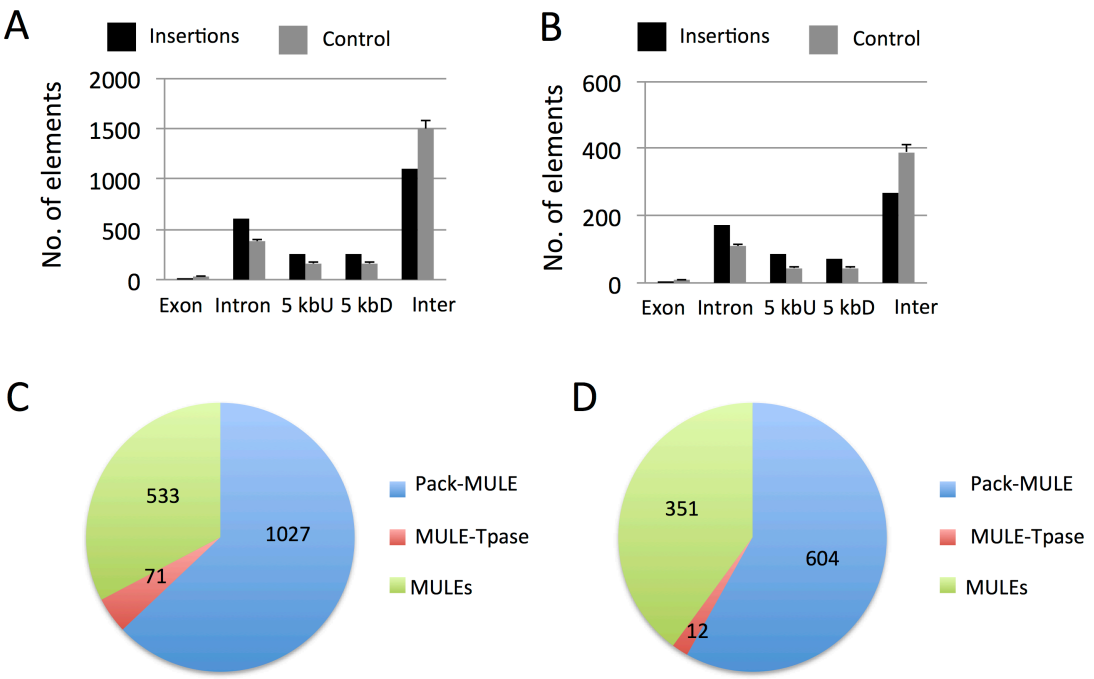

Figure 3-2. Distribution of *A. aegypti* MULE insertion sites and different classes of MULEs..

Number of long-TIR MULE (A) and short-TIR MULE (B) insertion sites in the *A. aegypti* genome and control data set is shown in black and grey, respectively. Mean ± s.d., n=1,000 (for control).

Copy number of each MULE classes of the long-TIR MULE (C) and short-TIR MULE (D) are shown. Pack-MULEs contain gene fragments; MULE-Tpase contain sequences with homology to known MULE transposases; and MULEs refer to elements without any identifiable host sequence.

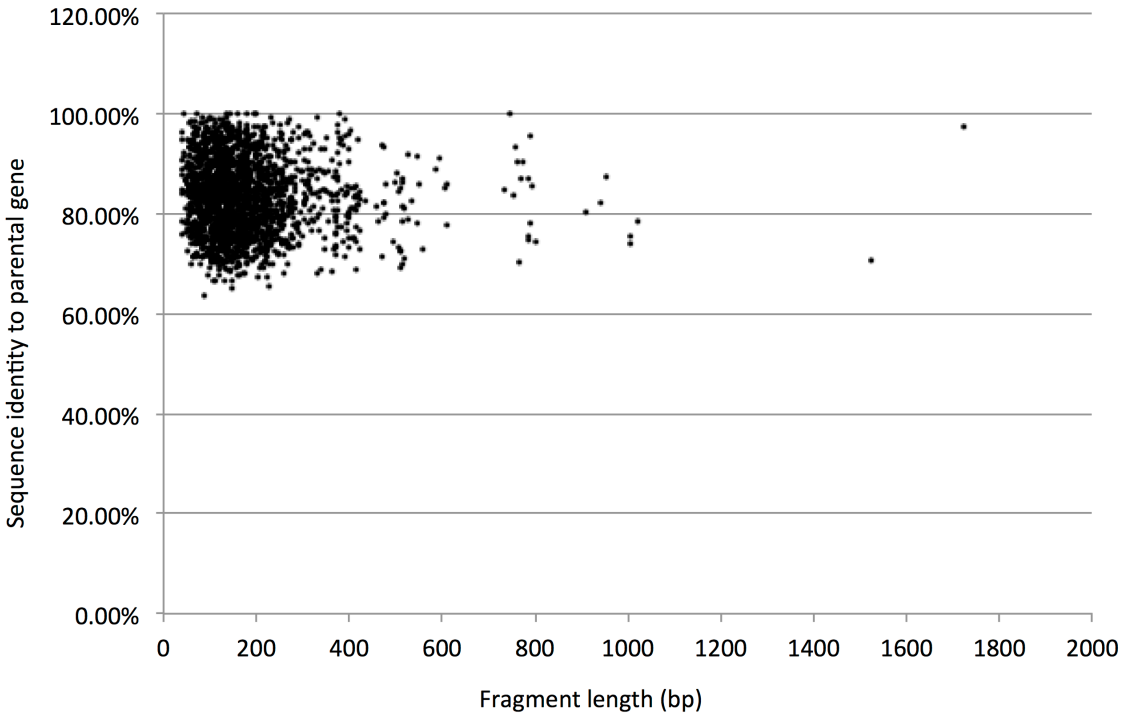

Figure 3-3. Size and sequence identity distribution of acquired fragments.

The size and sequence identity to host genes of 2,003 acquired fragments are shown.

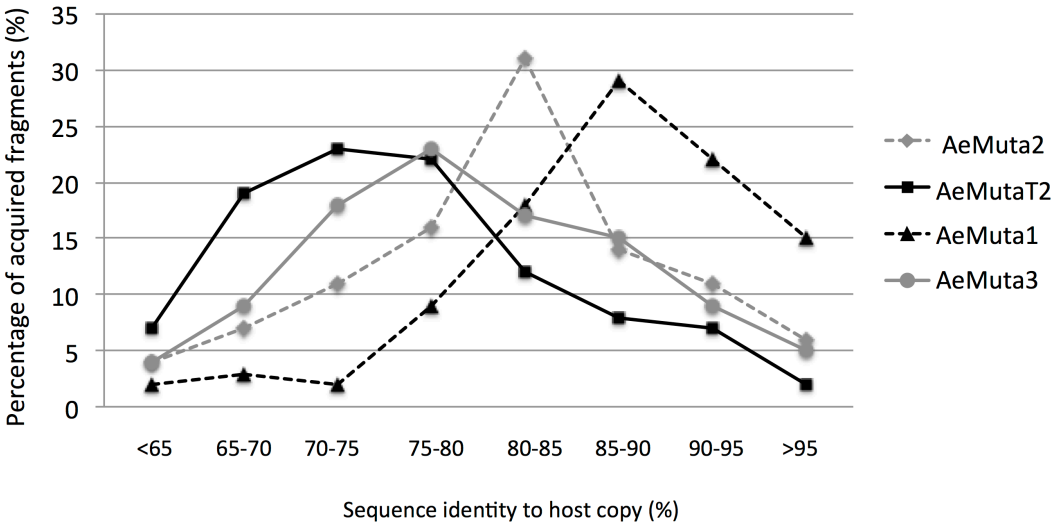

Figure 3-4. Estimated divergence of four Pack-MULE families.

Sequence identity between Pack-MULE acquired fragments and their host copies are shown. Four Pack-MULE families (*AeMuta2, AeMutaT2, AeMuta1*, and *AeMuta3*) are analyzed.

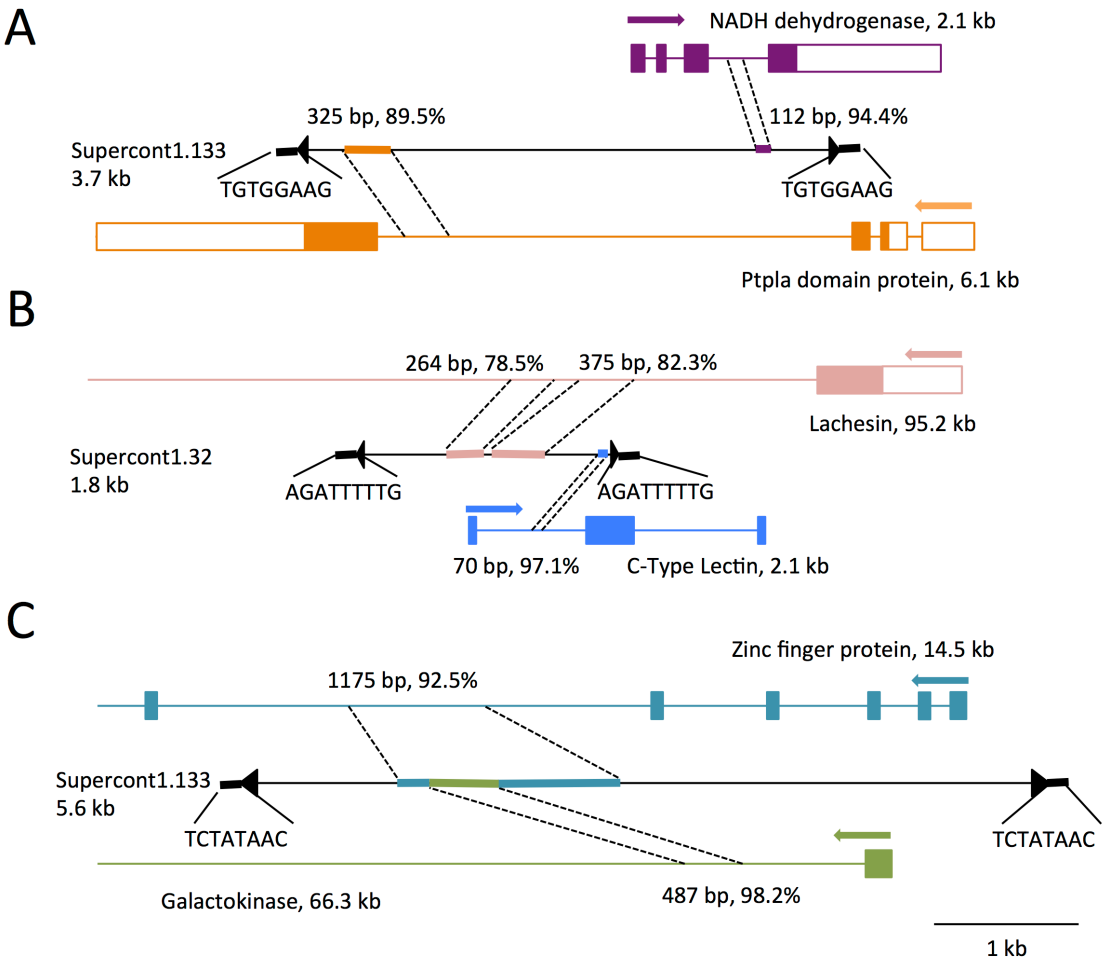
A
NADH dehydrogenase, 2.1 kb
325 bp, 89.5%
112 bp, 94.4%
Supercont1.133
3.7 kb
TGTGGAAG
TGTGGAAG
Ptpla domain protein, 6.1 kb
B
264 bp, 78.5%
375 bp, 82.3%
Lachesin, 95.2 kb
Supercont1.32
1.8 kb
AGATTTTTG
AGATTTTTG
70 bp, 97.1%
C-Type Lectin, 2.1 kb
C
Zinc finger protein, 14.5 kb
1175 bp, 92.5%
Supercont1.133
5.6 kb
TCTATAAC
TCTATAAC
Galactokinase, 66.3 kb
487 bp, 98.2%
1 kb

Figure 3-5. Examples of transcribed Pack-MULEs.

Genomic location and length of each Pack-MULE is shown on the left, TIRs are shown as black arrowheads, TSDs are shown as black boxes (with sequences underneath), colored boxes indicate acquired fragments, dashed lines connect acquired fragments and the host sequences. In host genes, colored boxes indicate coding sequences, open boxes indicate UTRs, colored lines indicate introns, and colored arrows indicate the transcription orientation of host genes. Annotation and length of host genes are also shown.

(A) A 3.7 kb Pack-MULE with a 112 bp fragment from the intron of the NADH dehydrogenase and a 325 bp fragment from the intron of the Ptpla domain protein.

(B) A 1.8 kb Pack-MULE with a 264 bp and a 375 bp fragment from the intron of the lachesin gene and a 70 bp fragment from the intron of the C-Type lectin.

(C) A 5.6 kb Pack-MULE involves possible step-wise acquisition events. First a 1175 bp fragment from the intron of a zinc finger protein was captured, next a 487 bp fragment from the intron of the galactokinase was inserted into the first fragment.

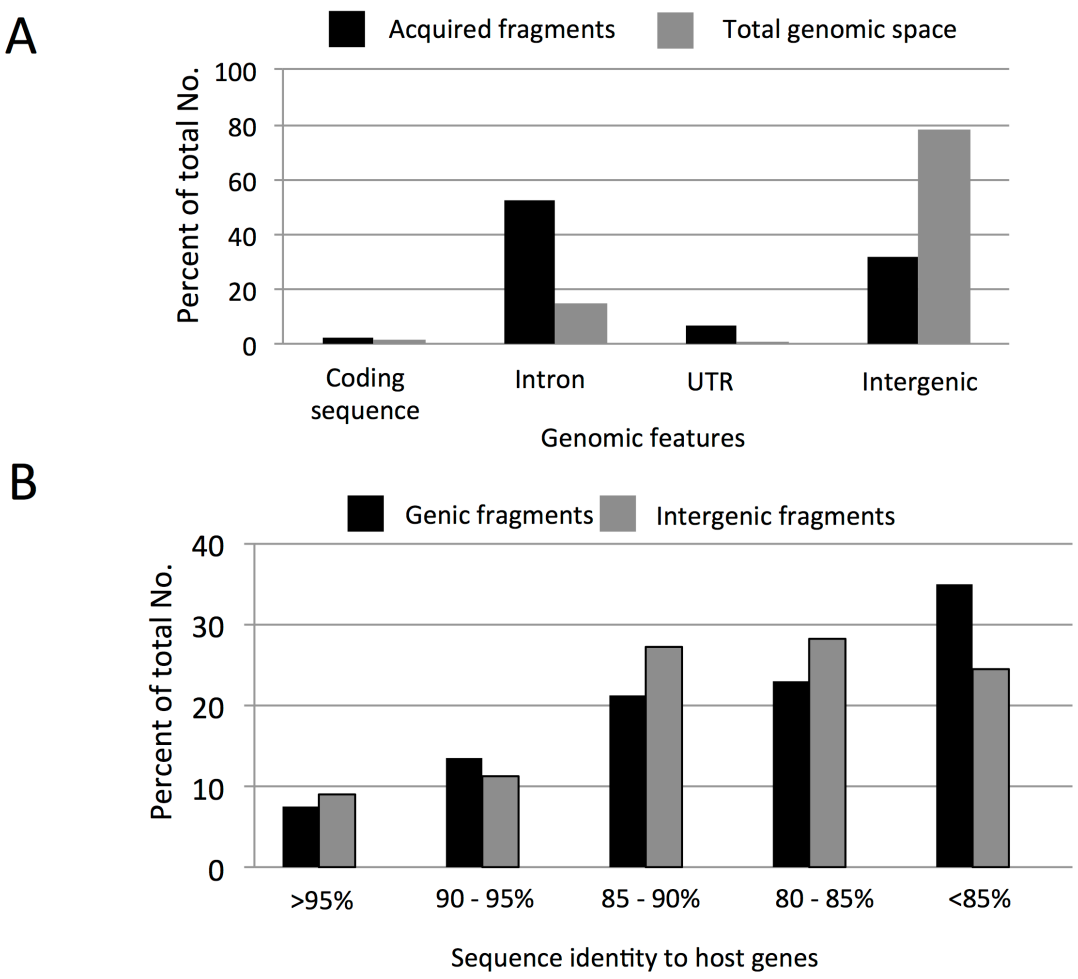

Figure 3-6. Analysis of the sequence origin of acquired fragments.

(A) Percentage of fragments acquired from each genomic feature is shown in black. Total genomic space of each genomic feature is shown in grey.

(B) Pack-MULE acquired fragments are classified by the sequence identity with host sequences. In each sequence identity group, percentage of fragments acquired from genes (black) and intergenic regions (grey) are shown.

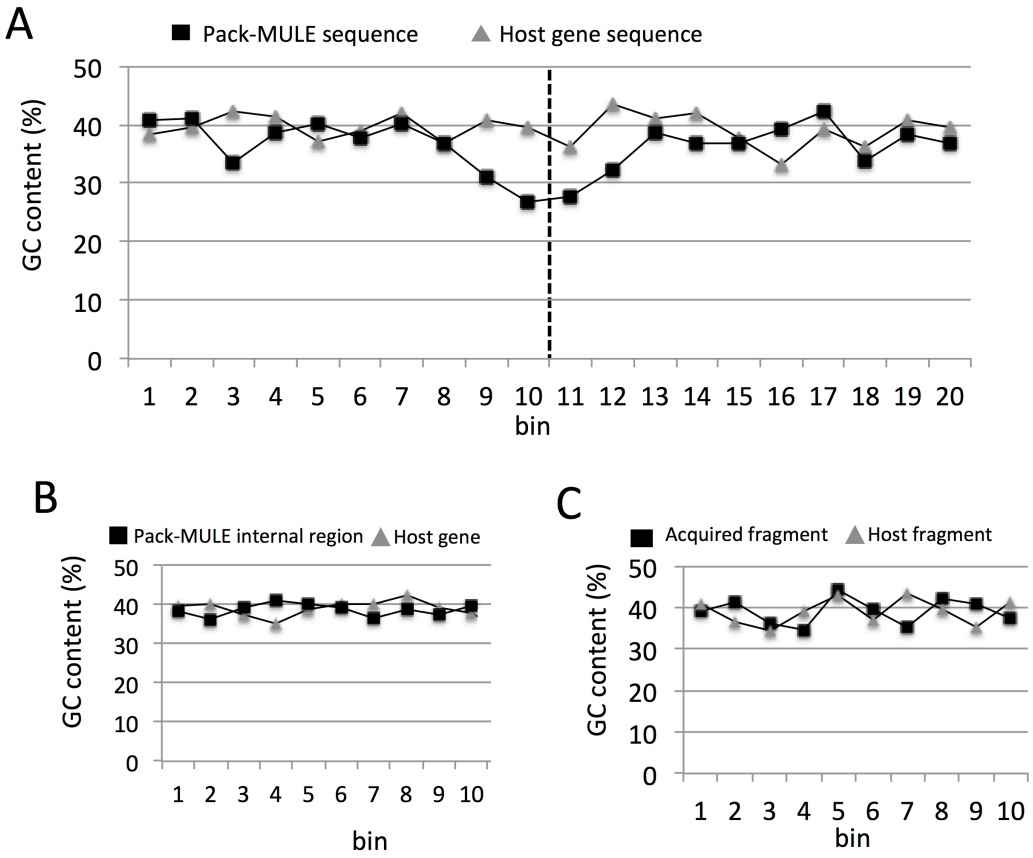

Figure 3-7. GC content of Pack-MULE and host genes flanking the acquisition junction.

(A) GC content Pack-MULE (black box) and host gene (grey triangle) sequences flanking the acquired fragment (or host copy). Each bin represents a 10 bp region, bin 1-10: 100 bp upstream flanking sequence; bin 11-20: 100 bp downstream flanking sequences. Dashed line indicates the position of the acquired fragment and host copy.
(B) GC content of the 175 Pack-MULE internal regions (black box) and corresponding host genes (grey triangle) used in this analysis. Each Pack-MULE or gene is divided into 10 equal-size bins.
(C) GC content of the 175 acquired fragments (black box) and corresponding host fragments (grey triangle) used in this analysis. Each fragment is divided into 10 equal-size bins.

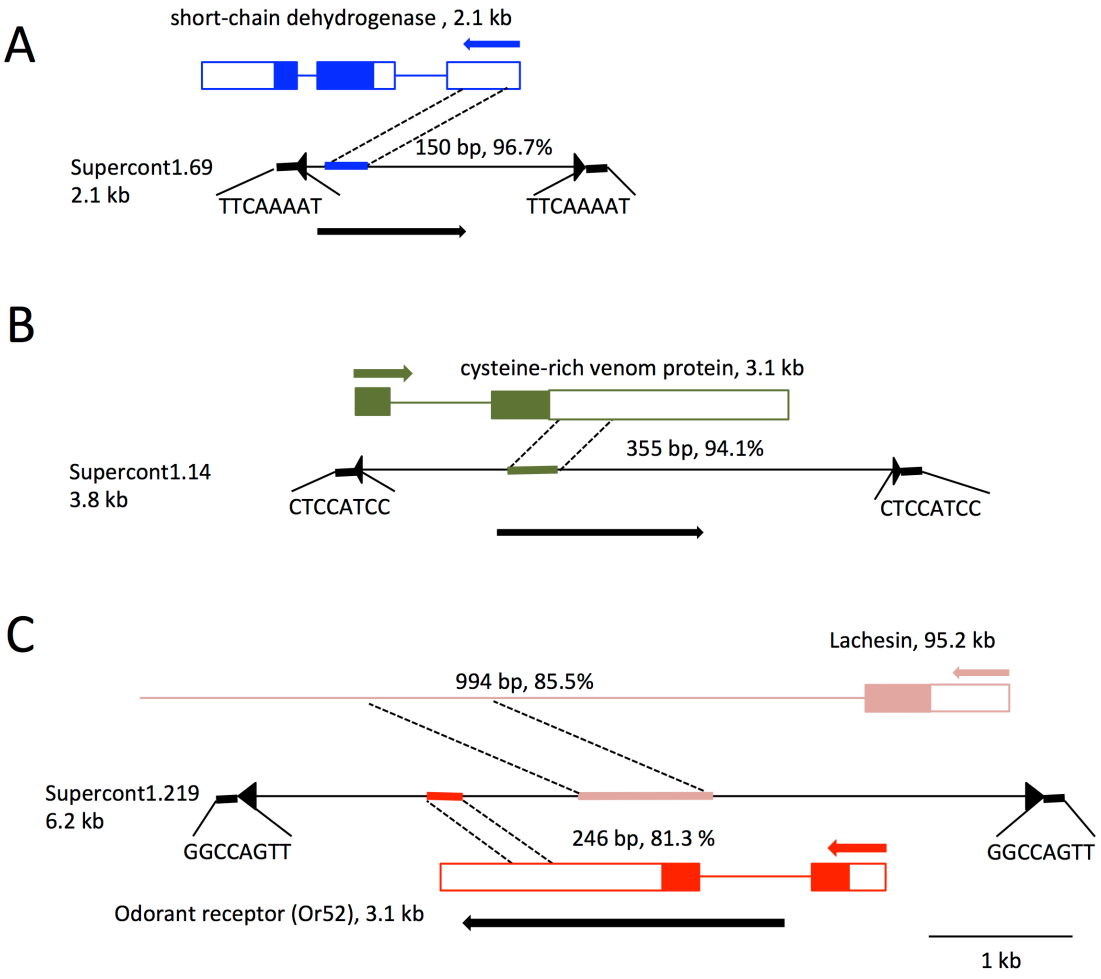
A
short-chain dehydrogenase , 2.1 kb
150 bp, 96.7%
Supercont1.69
2.1 kb
TTCAAAAT
TTCAAAAT
B
cysteine-rich venom protein, 3.1 kb
355 bp, 94.1%
Supercont1.14
3.8 kb
CTCCATCC
CTCCATCC
C
Lachesin, 95.2 kb
994 bp, 85.5%
Supercont1.219
6.2 kb
GGCCAGTT
246 bp, 81.3 %
GGCCAGTT
Odorant receptor (Or52), 3.1 kb
1 kb

Figure 3-8. Examples of transcribed Pack-MULEs.

Genomic location and length of each Pack-MULE is shown on the left, TIRs are shown as black arrowheads, TSDs are shown as black boxes (with sequences underneath), colored boxes indicate acquired fragments, dashed lines connect acquired fragments and the host sequences. In host genes, colored boxes indicate coding sequences, open boxes indicate UTRs, colored lines indicate introns, and colored arrows indicate the transcription orientation of host genes. Black arrows indicate the size, orientation, and position of the Pack-MULE transcript. Annotation and length of host genes are also shown.

(A) An 8 copy Pack-MULE with a 302 bp fragment from the short-chain dehydrogenase, it produces a 1.1 kb transcript.

(B) A 3.8 kb Pack-MULE with a 55 bp and a 375 bp fragment from the cysteine-rich venom protein, it produces a 1.4 kb transcript.

(C) A 6.2 kb Pack-MULE with two fragments from the odorant receptor (Or52) and the lachesin gene, it produces a 2.2 kb transcript.

Table 3-1. Number and length of multi-copy Pack-MULE subfamilies.

Copy No.	1	2	3	4	5	6	8	17
Subfamily No.	989	27	9	5	3	1	1	1
Ave. element length (bp)	2970	3119	2992	2911	2394	2038	2099	1644
Ave. fragment No.	1.8	1.6	2.2	1.8	1.7	2	1	2
Ave. fragment length (bp)	175	194	218	253	151	176	302	79
Ave. sequence identity (%) *	82.8	85.1	83.9	88.2	90.6	79.3	91.9	81.2

* The average sequence identity between acquired fragments and the corresponding parental gene.

Table 3-2. Comparison of *A. aegypti* and rice Pack-MULEs.

	A. Aegypti	rice
No. of MULEs (with TSD)	2,235	13,857
No. of Pack-MULEs	1136 (50.8% *)	2,924 (21.2% *)
No. of chimeric Pack-MULEs	5,40 (47.5% **)	656 (22.4% **)
No. of parental genes	410	1,557
No. of fragment acquired from coding sequences	66 (3.3% ***)	2156 (66.3% ***)
No. of transcribed Pack-MULEs	124 (10.9%)	613 (21.0%)
No. of translated Pack-MULEs	0	28 (0.96%)
No. of Pack-MULEs with self-sRNA ****	176 (15.5%)	1,722 (58.9%)
No. of Pack-MULEs with shared-sRNA ****	205 (18.1%)	2,168 (74.1%)

* Fraction of MULEs that are Pack-MULEs.

** Fraction of Pack-MULEs that are chimeric elements.

*** Fraction of gene fragments that are acquired from coding sequences.

**** sRNAs associated with internal region of Pack-MULEs

Conclusions

Mutator-like transposable element (MULE) is an important TE superfamily with members in fungi, plants, and animals. Prior studies showed that their movement and amplification alter the genome through a number of mechanisms including insertions, excisions, duplications and translocations. Gene acquisition by Pack-MULEs may also impact the host genome by altering gene expression and creating genetic novelty. However, despite their importance and abundance,, the mechanism of MULE transposition and gene acquisition is unknown.

In Chapter 1, we identified the first active animal MULE, from the mosquito, *A. aegypti*. Our data indicates that the *Muta1* transposase can efficiently catalyze both excision and reinsertion reactions in yeast despite the significant evolutionary distance. Mutagenesis analysis reveals that several conserved amino acids, including the DDE triad, play important roles in *Muta1* transposase function. In addition, presence of a donor site target site duplication (TSD) also impacts transposition frequency and quality. This is the first report of the transposition of a non-plant MULE in a heterologous system and provides the first experimental evidence for the functional significance of the DDE domain in the transposition of MULEs. With characteristics such as high transposition activity, precise excision, and no target sequence preference, *Muta1* could be crafted into an effective tool for forward mutation analyses in mosquitoes.

In Chapter 2, we dissected the pathway underling the DNA breakage and joining reactions of *Muta1* using *in vitro* and yeast transposition assays.. The transposition reaction involves double-strand break with hairpin formation in flanking DNA and 3' OH joining to the target DNA. Similar transposition mechanisms used by MULEs, *hAT* elements and the V(D)J recombination system provides new insights to understand their evolutionary relationships. Additionally, biochemical and yeast assays revealed the involvement of the subterminal repeats of foldback transposons in transposition.

Our study reveals that *A. aegypti* Pack-MULE mediate the duplication of host gene fragments on a large scale, including 1,378 Pack-MULEs, 2,249 acquired gene fragments and 423 host genes, this is the first report of Pack-MULEs in non-plant species. *A. aegypti* Pack-MULEs preferentially insert into gene rich regions and prefer to acquire sequences from genic regions. Although only a small proportion of *A. aegypti* Pack-MULEs are expressed and associated with sRNAs, they are involved in the acquisition, amplification and rearrangement of host gene fragments on a large scale, and providing a mechanism for shuffling the genetic material in *A. aegypti.*

In summary, our work identified the Muta1 element, confirmed its transposition activity in a yeast assay; dissected the transposition mechanism using in vitro assays, and provided experimental evidence to support the close evolutionary

relationship reported previously between the MULE and *hAT* superfamilies. Our analysis also revealed that the gene acquisition and rearrangement of *A. aegypti* Pack-MULEs provide a means for the genome evolution.

References

Agrawal, A., Eastman, O. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 1998, 394, 744–751.

Aguiar ERGR, Olmo RP, Paro S, Ferreira FV, de Faria Isaque João da S, Todjro YMH, Lobo FP, Kroon EG, Meignin C, Gatherer D, Imler JL, Marques JT: Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. Nucleic Acids Research 2015, 43(13): 6191–6206.

Akbari OS, Antoshechkin I, Amrhein H, Williams B, Diloreto R, Sandler J, Hay BA: The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. G3 (Bethesda) 2013, 3:1493-1509.

Alberti S, Gitler AD, Lindquist S: A suite of gateway (R) cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. Yeast 2007, 24:913-919.

Almeras L, Fontaine A, Belghazi M, Bourdon S, Boucomont-Chapeaublanc E, Orlandi-Pradines E, Baragatti M, Corre-Catelin N, Reiter P, Pradines B, Fusai T, Rogier C: Salivary gland protein repertoire from Aedes aegypti mosquitoes. Vector Borne Zoonotic Dis 2010, 10:391–402.

Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW: The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. BMC Genomics 2011, 12:606.

Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F: Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science 2010, 330:86-88.

Babu MM, Iyer LM, Balaji S, Aravind L: The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Res 2006, 34(22): 6505-20.

Babu MM, Iyer LM, Balaji S, Aravind L: The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. Nucleic Acids Res 2006, 34(22): 6505-20.

Bao W, Jurka MG, Kapitonov VV, Jurka J: New Superfamilies of Eukaryotic DNA Transposons and Their Internal Divisions. Mol Biol Evol 2009, 26(5): 983–993.

Bao W, Kojima KK, Kohany O: Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015, 6:11.

Bao, Z. & Eddy, S. R: Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 2002,12:1269–1276.

Bayyareddy K, Zhu X, Orlando R, Adang MJ: Proteome Analysis of Cry4Ba Toxin-interacting Aedes aegypti Lipid Rafts using geLC-MS/MS. J Proteome Res 2012, 11 (12):5843-5855.

Bennetzen JL, Springer PS: The generation of Mutator transposable element subfamilies in maize. Theor Appl Genet 1994, 87:657–667.

Bennetzen JL, Swanson J, Taylor WC, Freeling M: DNA insertion in the first intron of maize Adh1 affects message levels: cloning of progenitor and mutant Adh1 alleles. Proc Natl Acad Sci USA 1984, 81:4125–4128.

Biedler J, Qi Y, Holligan D, della Torre A, Wessler SR, Tu Z: Transposable element (TE) display and rapid detection of TE insertion polymorphism in the Anopheles gambiae species complex. 2003, 12(3): 211-6.

Bingham, PE, Zachar Z: Retrotransposons and the FB transposon from Drosophila melanogaster. in Mobile DNA 1989, pp. 485–502.

Brillet B, Bigot Y, Auge-Gouillou C: Assembly of the Tc1 and mariner transposition initiation complexes depends on the origins of their transposase DNA binding domains. Genetica 2007, 130(2):105–120.

Bureau TE, White SE, Wessler SR: 1994. Transduction of a cellular gene by a plant retroelement. Cell 77: 479–480.

Carels N, Bernardi G: Two classes of genes in plants. Genetics 2000, 154:1819–1825.
Chalvet F, Grimaldi C, Kaper F, Langin T, Daboussi MJ: Hop, an active Mutator-like element in the genome of the fungus Fusarium oxysporum. Mol. Biol. Evol 2003, 20:1362–1375.

Chen XG, et al: Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci USA 2015, 112:5907–5915.

Colot V, Haedens V, Rossignol JL: Extensive, nonrandom diversity of excision footprints generated by Ds-like transposon ascot-1 suggests new parallels with V(D)J recombination. Mol Cell Biol 1998, 18:4337–4346.

Craig, N., Craigie, R., Gellert, M. & Lambowitz, A. Mobile DNA II (ASM, Washington DC, 2002).

Eisen JA, Benito MI, Walbot V: Sequence similarity of putative transposases links the maize Mutator autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res 1994, 22:2634–2636.

Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J: High-frequency P element loss in Drosophila is homolog dependent. Cell 1990;62:515–25.

Feng X, Colloms SD. In vitro transposition of ISY100, a bacterial insertion sequence belonging to the Tc1/mariner family. Mol Microbiol 2007; 65:1432–1443.

Ferguson AA, Zhao D, Jiang N: Selective acquisition and retention of genomic sequences by Pack-Mutator-like elements based on guanine-cytosine content and the breadth of expression. Plant Physiol 2013, 163:1419–1432.

Feschotte C, Pritham EJ: DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 2007, 41:331–368.

Finnegan DJ: Eukaryotic transposable elements and genome evolution.Trends Genet 1989, 5:103–107.

Fraser MJ: Insect Transgenesis: Current Applications and Future Prospects. Annu Rev Entomol 2012, 57:267–89.

Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem 2002, 71:101–132.

Han YJ, Burnette JM, Wessler SR: TARGeT: A web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. Nucleic Acids Res 2009, 37:e78.

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu SH, Jiang N: The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell 2009, 21:25–38.

Hancock CN, Zhang F, Wessler SR: Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. Mob DNA 2010, 1:5.

Hencken C, Li X, Craig NL: Functional characterization of an active Rag-like transposase. Nat Struct Mol Biol 2012, 19(8):834-6.

Hennig S, Ziebuhr W: A transposase-independent mechanism gives rise to precise excision of IS256 from insertion sites in Staphylococcus epidermidis. J Bacteriol 2008, 190:1488-1490.

Hennig, S, Ziebuhr, W: A transposase-independent mechanism gives rise to precise excision of IS256 from insertion sites in Staphylococcus epidermidis. J Bacteriol 2008, 190:1488–1490.

Hickman AB, Chandler M, Dyda F: Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit Rev Biochem Mol Biol 2010, 45:50–69.

Hickman AB, Ewis HE, Li X, Knapp JA, Laver T, Doss AL, Tolun G, Steven AC, Grishaev A, Bax A, Atkinson PW, Craig NL, Dyda F: Structural basis of hAT transposon end recognition by Hermes, an octameric DNA transposase from Musca domestica. Cell 2014, 158:353–367.

Hiom K, Melek M, Gellert: DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 1998; 94:463–470.

Holt RA, et al: The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002, 298:129–149.

Hugo LE, Monkman J, Dave KA, Wockner LF, Birrell GW, Norris EL, Kienzle VJ, Sikulu MT, Ryan PA, Gorman JJ, Kay BH: Proteomic biomarkers for ageing the mosquito Aedes aegypti to determine risk of pathogen transmission. PLoS ONE 2013, 8: e58656-10.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR: Pack-MULE transposable elements mediate gene evolution in plants. Nature 2004, 431:569–573.

Jiang N, Ferguson AA, Slotkin RK, Lisch D: Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc Natl Acad Sci USA 2011, 108:1537-1542.

Jiang, N, Bao, Z, Zhang, X, Eddy SR, Wessler SR: Pack-MULE transposable elements mediate gene evolution in plants. Nature 2004, 431:569–573.

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE: The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 2005, 15:1292–1297.

Kapitonov VV, Jurka J: Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 2001 98: 8714–8719.
Kapitonov VV, and Jurka J: Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 2001, 98:8714–8719.

Kazazian Jr HH: Mobile elements: drivers of genome evolution. Science 2004, 303:1626–1632.

Kim MS, Lapkouski M, Yang W, Gellert M: Crystal structure of the V(D)J recombinase RAG1-RAG2. Nature 2015, 518, 507–511.

Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9:357–360.
Le QH, Wright S, Yu Z, Bureau TE: Transposon diversity in Arabidopsis thaliana. Proc Natl Acad Sci 2000, 97:7376–7381.

Li X, Harrell RA, Handler AM, Beam T, Hennessy K, Fraser Jr MJ (2005) piggyBac internal sequences are necessary for efficient transformation of target genomes. Insect Mol Biol 14: 17–30.

Li Y, Harris L, Dooner HK: TED, an autonomous and rare maize transposon of the Mutator superfamily with a high gametophytic excision frequency. Plant Cell 2013, 25:3251–3265.

Lisch, D. Mutator transposons. Trends Plant Sci, 2002, 7: 498–504.

Lisch D, Jiang N. Mutator and MULE transposons. Handbook Maize 2009, 3:277–306.

Lish D: Regulation of the Mutator System of Transposons in Maize. Methods in Molecular Biology 2013, 1057:123-142.

Liu D, Mack A, Wang R, Galli M, Belk J, Ketpura NI, Crawford NM: Functional dissection of the cis-acting sequences of the Arabidopsis transposable element Tag1 reveals dissimilar subterminal sequence and minimal spacing requirements for transposition. Genetics 2000, 157: 817–830.

Loessner I, Dietrich K, Dittrich D, Hacker J, Ziebuhr W: Transposase-dependent formation of circular IS256 derivatives in Staphylococcus epidermidis and Staphylococcus aureus. J Bacteriol 2002, 184:4709–4714.

Lopes FR, Silva JC, Benchimol M, Costa GG, Pereira GA, Carareto CM: The protist Trichomonas vaginalis harbors multiple lineages of transcriptionally active Mutator-like elements. BMC Genomics 2009, 10:330.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK: A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci USA 2008, 8105(27):9272-7.

Ma Y, Pannicke U, Schwarz K, Lieber MR: Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell 2002, 108:781–794.

Marchette NJ, Garcia R, Rudnick A: Isolation of Zika virus from Aedes aegypti mosquitoes in Malaysia. Am J Trop Med Hyg 1969, 18:411–5.

Marinotti O, Ngo T, Kojin BB, Chou SP, Nguyen B, Juhn J, Carballar-Lejarazú R, Marinotti PN, Jiang X, Walter MF, TU Z, Gershon PD, James AA: Integrated proteomic and transcriptomic analysis of the Aedes aegypti eggshell. BMC Dev Biol 2014, 14: 15.

Marquez CP, Pritham EJ: Phantom, a new subclass of Mutator DNA transposons found in insect viruses and widely distributed in animals. Genetics 2010, 185(4):1507–1517.

Marzo M, Liu D, Ruiz A, Chalmers R: Identification of multiple binding sites for the THAP domain of the Galileo transposase in the long terminal inverted-repeats. Gene 2013, 525:84-91.

Mitra R, Fain-Thornton J, Craig NL: piggyBac can bypass DNA synthesis during cut and paste transposition. Embo J. 2008, 27: 1097-1109.

Moran JV, DeBerardinis RJ, Kazazian HH: Exon shuffling by L1 retrotransposition. Science 1999, 283:1530–1534.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. 2005, 37: 997–1002.

Neafsey DE, et al: Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science 2015, 347:1258522.

Nene V, et al: Genome sequence of Aedes aegypti, a major arbovirus vector. Science 2007, 316:1718–1723.

Nesmelova IV, Hackett PB: DDE transposases: Structural similarity and diversity. Adv Drug Deliv Rev 2010, 62(12): 1187-95.

Neuveglise C, Chalvet F, Wincker P, Gailardin C, Casaregola S: Mutator-like element in the yeast Yarrowia lipolytica displays multiple alternative splicings. Eukaryot Cell 2005, 4:615–624.

Nunes, AT, Brito, NF, Oliveira DS, Araujo GD, Nogueira FC, Domont GB, Moreira MF, Moreira LM, Soares MR, Melo AC: Comparative proteome analysis reveals that blood and sugar meals induce differential protein expression in Aedes aegypti female heads. Proteomics 2016, 16(19):2582-2586.

O'Brochta DA, Sethuraman N, Wilson R, Hice RH, Pinkerton AC, Levesque CS, Bideshi DK, Jasinskiene N, Coates CJ, James AA, Lehane MJ, Atkinson PW: Gene vector and transposable element behavior in mosquitoes. J Exp Biol 2003, 206:3823–3834.

Oktarianti R, Senjarini K, Hayano T, Fatchiyah F, Aulanni'am: Proteomic analysis of immunogenic proteins from salivary glands of Aedes aegypti. J Infect Public Health 2015, 8(6):575-82

Parker JS, Roe SM, Barford D: Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. EMBO J 2004, 23:4727–4737.

Pickeral OK, Makalowski W, Boguski MS, Boeke JD: Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res 2000, 10: 411–415.
Pingoud A, A. Jeltsch: Recognition and cleavage of DNA by type-II restriction endonucleases. Eur J Biochem 1997, 246:1–22.

Plasterk RH. The origin of footprints of the Tc1 transposon of Caenorhabditis elegans. EMBO J 1991, 10:1919–1925.

Plasterk RHA, Izsvák Z, Ivics Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 1999, 15:326–332.

Popova-Butler A, Dean DH: Proteomic analysis of the mosquito Aedes aegypti midgut brush border membrane vesicles. J Insect Physiol 2009, 55:264–272.

Potter S, Truett M, Phillips M, Maher A: Eucaryotic transposable elements with inverted terminal repeats. Cell 1980, 17:429–439.

Pritham EJ, Feschotte C, Wessler SR: Un-expected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol Biol Evoanl 2005, 22:1751–1763.

Rebatchouk D, Narita JO: Foldback transposable elements in plants. Plant Mol Biol 1997, 34:831–835.

Robertson DS: Characterization of a mutator system in maize. Mutat Res 1978, 51:21–28.
Sakai J, Chalmers RM, Kleckner N: Identification and characterization of a pre-cleavage synaptic complex that is an early intermediate in Tn10 transposition. EMBO J 1995, 14:4374-4383.

Scali C, Nolan T, Sharakhov I, Sharakhova M, Crisanti A, Catteruccia F: Post-integration behavior of a Minos transposon in the malaria mosquito Anopheles stephensi. Mol Genet Genomics 2007, 278:575–584.

Sirot LK, Poulson RL, McKenna MC, Girnary H, Wolfner MF, Harrington LC: Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, Aedes aegypti: potential tools for control of female feeding and reproduction. Insect Biochem Mol Biol 2003, 38:176–189.

Song JJ, Smith SK, Hannon GJ, Joshua-Tor L. Crystal structure of Argonaute and its implications for RISC slicer activity. Science 2004, 305:1434–1437.

Sutton WD, Gerlach WL, Schwartz D, Peacock WJ: Molecular analysis of Ds controlling element mutations at the Adhl locus of maize. Science 1983, 223:1265-1268.

Talbert LE, Chandler VL: Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol. Biol. Evol. 1988, 5:519–529.

Turcotte K, Srinivasan S, Bureau TE: Survey of transposable elements from rice genomic sequences. Plant J. 2001, 25:169–179.

Walbot V, and Rudenko GN: MuDR/Mu transposable elements of maize. In Mobile DNA II, (Washington, DC: America Society of Microbiology Press), pp. 533–564.

Weil CF, Kunze R: Transposition of maize Ac/Ds transposable elements in the yeast Saccharomyces cerevisiae. Nature Genet 2000, 26:187–190.

Weil CF, Kunze R: Transposition ofmaize Ac/Ds transposable elements in the yeast Saccharomyces cerevisiae. Nature Genet 2000, 26:187–190.

Wicker T, Sabo F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaid O, Paux E, SanMiguel P, Schulman AH: A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007, 8:973–982.

Windsor AJ, Waddell CS: FARE, a new family of foldback transposons in Arabidopsis. Genetics 2000, 156:1983–1995.

Womack M: The yellow fever mosquito, Aedes aegypti. Wing Beats 1993, 5(4):4.

Wong GKS, Wang J, Tao L, Tan J, Zhang J, Passey DA, YU J: Compositional gradients in Gramineae genes. Genome Res 2002, 12:851–856.

Xu Z, Yan X, Maurais S, Fu H, O'Brien DG, Mottinger J, Dooner HK: Jittery, a Mutator distant relative with a paradoxical mobile behavior: Excision without reinsertion. Plant Cell 2004, 16:1105–1114.
Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y: Resistance to gap repair of the transposon Tam3 in Antirrhinum majus: a role of the end regions. Genetics 1999, 153:1899–1908

Yang G, Weil CF, Wessler SR: A rice Tc1/mariner-like element transposes in yeast. Plant Cell 2006, 18:2469–78.

Yang N, Kazazian HH: L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nature Struct. Mol. Biol 2006, 13:763–771.

Yu Z, Wright SI, Bureau TE: Mutator-like elements in Arabidopsis thaliana: Structure, diversity and evolution. Genetics 2000,156:2019–2031.

Yuan YW, Wessler SR: The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci USA 2011, 108:7884–7889.

Zabala G, Vodkin LO: The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 2005, 17:2619–2632.

Zhao D, Ferguson A, Jiang N: Transposition of a rice Mutator-like element in the yeast Saccharomyces cerevisiae. Plant Cell 2015, 27:132–148.

Zhou L, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL: Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 2004, 432:995–1001.

Zhou L, Mitra R, Hickman AB, Dyda F, Craig NL: Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 2004, 432:995–1001.

Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D: Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics 2009, 10:47.